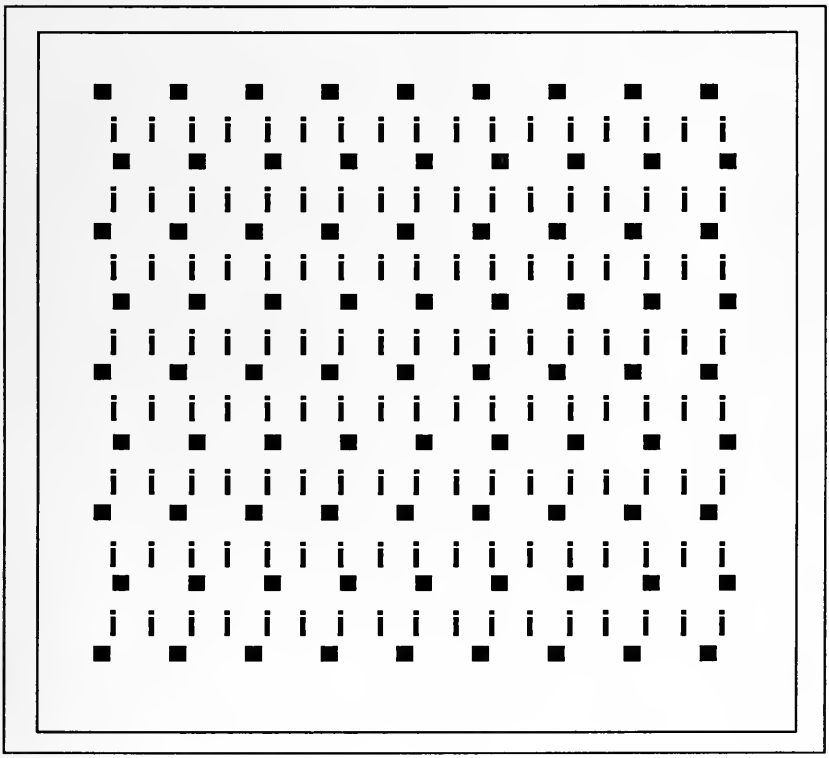
IASSIST

QUARTERLY

VOLUME 14 Fall/Winter 1990 NUMBER 3/4



IASSIST QUARTERLY

The IASSIST QUARTERLY represents an international cooperative effort on the part of individuals managing, operating, or using machine-readable data archives, data libraries, and data services. The QUARTERLY reports on activities related to the production, acquisition, preservation, processing, distribution, and use of machine-readable data carried out by its members and others in the international social science community. Your contributions and suggestions for topics of interest are welcomed. The views set forth by authors of articles contained in this publication are not necessarily those of IASSIST.

Information for Authors

The QUARTERLY is published four times per year. Articles and other information should be typewritten and double-spaced. Each page of the manuscript should be numbered. The first page should contain the article title, author's name, affiliation, address to which correspondence may be sent, and telephone number. Footnotes and bibliographic citations should be consistent in style, preferably following a standard authority such as the University of Chicago press *Manual of Style* or Kate L. Turabian's *Manual for Writers*. Where appropriate, machine-readable data files should be cited with bibliographic citations consistent in style with Dodd, Sue A. "Bibliographic references for numeric social science data files: suggested guidelines". *Journal of the American Society for Information Science 30(2):77-82, March 1979.* If the contribution is an announcement of a conference, training session, or the like, the text should include a mailing address and a telephone number for the director of the event or for the organization sponsoring the event. Book notices and reviews should not exceed two double-spaced pages. Deadlines for submitting articles are six weeks before publication. Manuscripts should be sent in duplicate to the Editor: Walter Piovesan, Research Data Library, W.A.C. Bennett Library, Simon Fraser University, Burnaby, B.C., V5A 1S6 CANADA. (604) 291-4349 E-Mail: USERDLIB@SFU.BITNET Book reviews should be submitted in duplicate to the Book Review Editor: Daniel Tsang, Main Library, University of California P.O. Box 19557, Irvine, California 92713 USA. (714) 856-4978 E-Mail: DTSANG@ORION.CF.UCI.EDU

Title: **Newsletter - International Association for Social Science Information Service and Technology**
ISSN - United States: 0739-1137

CONTENTS

Volume 14 Number 3/4 Fall/Winter 1990

FEATURES

News Beat

Data Bank on the U.S.A. and Soviet-American Relations

by Tatyana Yudina [1]
Senior Researcher, SPASIBO
Institute of the USA and Canada Studies USSR,
Academy of Sciences Moscow, USSR

Introduction

Our Institute - Institute of USA and Canada Studies of the USSR Academy of Sciences - was established in 1969. There are several departments in it; USA domestic policy, foreign policy, military policy, agriculture, management system, and others. Some 200 researchers and post graduates work in our institute.

Two years ago several dozen personal computers (IBM - AT and XT clones) were purchased and a new section was established - that of applied research and informatics. Invited to join the section were researchers interested in new technologies and new sources of information. I was one of 12 memebers to join this new group. We also have two professional full-time programmers and several part-time programmers working with us; not enough to meet our needs, however. In this we are not alone as I learned from my attendance of the 1990 IASSIST conference. I discovered that this is a common problem among data libraries throughout the world - understaffed with the personnel that are needed most of all.

Our section has several major functions:

- teaching researchers computer skills
- studying the software market for emerging applicationsthat may be rquired by our department an providing traing in software use
- designing the computer-based information-retrieval system for our institute and coordinating the efforts of different departments of our institute taking into account the strategic aim of our data base - to have as much data as possible,
- studying and evaluating new sources of information; CD-ROMs, on-line data-banks, archives of machine-readable data etc.

These are our main tasks to say nothing about educating our chiefs and encouraging them to give money for ongoing operations and development of our programs.

We were among the first of the humanitarian institutes to use personal computers, with limited access to consultants, and with no prior computing experience, we have made do through trial and error.

Computer-based Information Retrieval System "USA and Soviet-American Relations

Our Institute is a section of the USA studies in the USSR, and our mandate is to serve as a national data bank on the USA and Soviet-American relations.

We are fortunate in that there is a great quantity and variaty of high quality data that the USA produces about itself. Our problem is one of finding funds for the purchase of the data we choose. Happily we have other responsibilties in addition to building data collections.

The system is multifunctional and multilingual, it contains data of various genres, original full texts or created especially for the system in Russian, in English and some in French. We hope that in future it will be supported by the automatic translation subsystem.

There are two blocks of information in our system - one about the USA and the other about the Soviet-American relations. Both blocks contain several types of data sets :

- the "USA reference module"; an electronic version of the USA Encyclopedia recently published by our institute,
- a statistical module consisting of long range time-series data on economic indicators in the USA,
- full-text documents, such as Soviet-American agreements from 1933 to present (updated on a fortnightly basis),
- a chronology of USA-USSR relations,
- articles and statistics about economic and trade relations between the two countries,
- mutual Soviet-American projects and joint ventures,
- biographies of the American political leaders, with a special module: "speeches" - short references of main ideas processed by our linguistic means (further they will be mentioned),
- archive collections,
- bibliography of publications about the USA and Soviet-American relations done by our library, which has started to create an on-line catalogue of its holdings.

In our data collection we have data sets on such subjects as: the Congress of the USA, President Bush's cabinet, White House personnel and the structure of the American administration, the constitution of USA, etc. These

subject data sets are being created by our researchers - we encourage such projects, since we simply can't afford to purchase such collections from the USA. In collectiong information our researchers make useof American online services such as Dialogue, Lexis-Nexis, among others, and which is funded by the Academy of Sciences.

Work organization

In developing our information system we have introduced some new elements of work organization - our researchers are invited to build data collections in their field of expertice. We did this for several reasons, the first one is that we hope that a researcher - expert in some particular sphere knows the information market better, can evaluate different sources, and choose the most informative ones. He is also supposed to process the data, to administer and update the collection. We encourage our researchers to start such data collections, though there are some social and psychological problems to this in that researchers become possive of their data. We hold competitions amoung the researchers, with the winners going to attend the summer workshops at the University of Michigan or Essex University , where they learn about new methods of analysis and modern software for personal computers. The idea of data collections built by experts is that they use the data as a base for their analysis while others mostly use it for information.

We especially encourage such data sets that may be used for modelling.

New Sources of Information

As for the on-line sources of data we have on-line communication with Soviet Press Agency which widely reports international news, and speeches of American political leaders. All the speeches of American political leaders which touche upon the problems of Soviet-American relations are supposed to be processed and added to our political "portraits" databases.

One of the most important source of data for our research is the data archive of the Inter-University Consortium for Political and Social Research at Univeristy of Michigan. We believe that soon we'll be honoured with the opportunity to join it and to access its data collections which will help us to raise the level of our research. For a decade we were unable to join ICPSR because of not fulfilling one of its major conditions - the USSR had no machine-readable information about itself in the market. By now the situation has changed - our State Statistics Agency (Goscomstat) announced that a special section has been organized with the aim to produce machine-readable collections of data about the development of the USSR.

Our other interest in joining the Inter-University Consortium for Political and Social research is its Summer School program This year we send two of our researchers to ICPSR's Summer Program and hope that it will help us to move to new levels of research using modern software and methodology in analyzing social and political proccesses.

Software

The software that is now being designed by our main programmer, Valentine Ponomarenko, is supposed to allow to use direct entry, scanners, on-line communication, and CD-ROMs to add data to our information systems. It will also provide us with an interface transition from data set to data set. Having information in several languages we come accross the problem of using English and cyrrilic alphabets which we manage to solve by having software designed by our prgrammers which is specific to our needs.

Linguistic means of analyses

Most of the documents we are dealing with are full-text and to create an effective information system is impossible without the built in and complex linguistic means of processing and analysing the full-text databases; thus a sub-system of advanced automatic analyses of the full text documents becomes an mandatory component.

In full-text documents special knowledge and information are accumulated in natural language. The idea of "restricted language", an approach adopted by most artificial intelligence systems, is irrelevant for the field of social and political sciences.

Moreover as for political text it's insignificant parts may be conveyed by so called key words. The more exact expression of the content calls for taking into account the relations between key words and the use of logical inference.

So we've started research in a field of structural linguistics which is absolutely untraditional for our institute. We have had a group of professional linguists assigned to our section, without costs to us, and are enthusiastically assisting our researchers in their tasks. The description of this project may be best addressed in a seperate article by Nina Leontieva, who is in charge of this work.

Financing

Our work is financed by the Academy of Sciences of the USSR. For the past two years we applied and received additional funds to work on our linguistic means of analysis project, to invite professional linguists and designers, and additional salaries for researchers working on this project. I realize how difficult it must have been for our academy chiefs to make the decision to finance

our idea of linguistic analysis and I am very pleased that they decided to finance the project. Our results to date are gratifying, we have developed several functional modules of linguistic analysis and other modules are in process - most notably the Russian Semantic Dictionary where political lexics are being fully described and the creation of the Thesaurus of Political Terms. Both of them may become commercial products and that is very important for us since it helps us to earn additional money to support our project.

Another funding source is the "USSR Congress of Peoples Deputies" database which was created by our institute. It is of interest to several universities in the USA and Europe. Version 1 contains the biographical information on deputies; education, profession, career; and information about the district the deputy represents, or public organization that elected him/her; the deputy's position in the Supreme Soviet and membership in committees and commissions. The database is in Russian. Names, geographic names, names of public organization have equivalents in English. We are planning to have Version 2 available soon which will have some additional fields - voting behavior of the deputies, voting results, etc.

So our section of applied research and informatics after two years of looking for our own way of developing new technologies and new thinking in our institute has some results,a lot of problems, and many ideas.

In conclusion I would like to express my great gratitude to all the IASSIST people who have made it possible for me to attend this conference - it was very helpful and informative. I hope that one day soon an IASSIST conference may be held in Moscow, USSR. IASSIST 90 has helped a lot to make this possible.

[1] A presentation to the IASSIST 90 Conference held in Poughkeepsie, N.Y. May 30 - June 2, 1990.

Career Achievement Award

On Febuary 21 the second IASSIST Career Achievment Award was awarded to Don Harrison. Judith Rowe presented the award to Don on behalf of our organization, at a retirement reception held in his office at the National Archives building in Washington, D.C. The award read, "IASSIST honors Donald Fisher Harrison for his career of dedicated service to the archival preservation of computerized information. February 1991." The authorization of the award followed the procedures approved last May by the General Assembly and finalized by the Admin Committee last October.

Tom Brown, IASSIST president and a colleague of Don, noted that "this marks a personal passing for me. For it was Don who introduced me to bits and bytes and everything else about machine readable data files sixteen years ago when I joined the staff of the National Archives. Many thanks Don!"

Iassist members join Tom Brown and the staff of the National Archives in extending the warmest of thanks to Don for his many contributions to IASSIST and his profession.

IASSIST

The Changing Nature of Networking in the Research Library Community

by Henriette D. Avram [1]
Associate Librarian for Collections Services
Library of Congress

Networking among libraries is certainly nothing new; to the contrary, libraries have long been pioneers in networking activities. Today, I want briefly to summarize that history, but in keeping with the theme of this session, my focus will be on future plans and prospects. We are in an exciting period now — one where the technology and our collective imagination are at such a confluence as to yield exhilarating results for libraries, especially the research library community.

From the beginning, networking among libraries has been propelled by our strongly held tradition of resource sharing. The 1960s and 70s witnessed libraries turning increasingly to the computer and the possibilities it held for automating library operations. The development of the MARC format at LC, the acceptance of the format by library practitioners, the adoption of the format structure as a national and international standard, and LC's distribution of its cataloging data in this format, marked the beginning of the true era of library networking as we define it today.

There then appeared organizations that were new on the scene, namely, bibliographic utilities. These organizations, created to serve the needs of libraries desirous of a central source for cataloging records at a reasonable cost, built growing files of cataloging data to which access was limited to institutions who were members of the particular utility. By the mid-1970s several large databases — OCLC, the Research Libraries Group's RLIN (Research Libraries Information Network), and WLN (the Western Library Network), along with the one at LC — coexisted in the United States, but could not be accessed and shared directly. To remedy this situation, efforts were initiated to enable libraries more easily to share data housed on dissimilar systems and thus was created the Linked Systems Project or LSP.

The International Organization for Standardization's Open Systems Interconnection (OSI) Reference Model was chosen by the LSP architects as the appropriate protocol package to run LSP because OSI would substantially reduce future development necessary to accommodate new systems, new applications, and new standards being developed in accordance with the OSI model.

Two LSP applications modules were developed — Record Transfer and Information Retrieval. Record Transfer enables records of any type and any number to be transported between systems. Information Retrieval permits users of one system to access a remote system and to view data found in the remote systems. Information Retrieval permits users of one system to access a remote system and to view data found in the remote system on their own system, invoking the familiar query command of their local system, effectively overcoming the problem of multiple syntaxes.

The OSI-based protocols for LSP were fashioned to be of general service and not application specific. Thus LSP applications can be expanded to other purposes, e.g., the Information Retrieval protocol, which is now an American National Standards Institute standard, is the basis of several projects being planned in the U.S. for accessing remote databases of all kinds, e.g., full text, abstract and indexing, and others.

Currently, LSP is being used to support the exchange of authority records. LC has distributed via LSP connections more than 2.5 million authority records to RLIN and over 1.5 to OCLC. LC has received via LSP, over 88,000 authority records created by RLIN and OCLC libraries which have been added to the file at LC and distributed via LC's Cataloging Distribution Service to libraries all over the world.

The next step is to support the exchange of bibliographic records which will enable records to be searched between systems, retrieved and then added to a particular database for use by its patrons. By this augmentation of LSP, a cooperative operation located at LC, NCCP (National Coordinated Cataloging Program), will gain increased efficiency. NCCP brings together eight research libraries that have agreed to contribute national-level cataloging records to the national database at LC for distribution to the Nation's libraries.

While the library community was availing itself of advances in technology to forge a national bibliographic network via LSP, using OSI protocols, other networks were evolving in the U.S. using different standards. The academic and scientific community was busily laying the foundation for a supernetwork supported by TCP/IP (Transmission Control Protocol/Internet Protocol) that

will support research and scientific investigations. This network, known as the Internet, connects universities across the U.S. and links them to supercomputer centers. The Internet is a long-haul network that provides national connectivity through the linking of regional networks which cover large geographic areas. NSFNet, the National Science Foundation Network, acts as the backbone of the Internet.

Because of the difference in standards used, the two networks being built were incompatible. The first step toward reconciling this incompatibility was to seek cooperation with EDUCOM (a consortium of U.S. institutions of higher learning). Accordingly, in 1987, Henriette Avram invited Ken King, President of EDUCOM, to come to LC for exploratory talks.

These initial meetings opened our eyes as librarians to the enormity of what was happening, what was being planned on the academic side, and also what was missing, i.e., much of what libraries were already doing or had already accomplished. The aim was to connect scholars' workstations on the Nation's campuses to each other as well as to supercomputer centers via the Internet to support research needs. Besides NSF, the players in this grand scheme were influential and represented big money interests — IBM, AT&T, and New York Telephone.

As more was learned of what was envisioned, an additional strong concern emerged — the research libraries on university campuses being wired to each other and to supercomputer centers were also part of the LSP environment. These libraries are the keepers and organizers of much of the information and data that feed the research process. It is the ability of libraries to organize information for retrieval and end user access that makes sharing data over networks viable. We should not lose sight of the importance of the organization of information and the critical role standards play in this process of organization. Indeed, it is this technical processing aspect that underpins and makes possible the research and reference functions that will become increasingly important as the supernetwork takes shape and the variety of data on it mushrooms.

EDUCOM has been instrumental in pushing and tracking legislation currently before Congress (S. 1067, sponsored by Senator Albert Gore). If approved, the evolution of the Internet will take on immense proportions. Of keen interest to research libraries is Title II of the bill, which calls for the creation of a high capacity National Research and Education Network (NREN), which will interconnect over 1,000 colleges, universities, research organizations, and, we hope, their libraries. Title II further specifies that — working with other agencies — by 1996, NSF would establish a multi-gigabit NREN capable of transmitting 100,000 typed pages or 1,000 satellite photographs in one second. The network is to be phased out when national, commercial high-speed networks can satisfy research needs.

Now that the networking infrastructure that will serve the Nation in its various components has been described, it is well worth spending a few minutes discussing the "content" of the Network, i.e., what kinds of information and data will be accessible over the Network. As I've tried to make clear, in terms of describing and formatting bibliographic data for efficient searching and retrieval, we're pretty much there. The standards are well defined and broadly applied within the library networking environment. For non-bibliographic data, however, we have some way to go yet. But strides are being made.

Having spent its first fourteen years focused primarily on networking of bibliographic data, the Library of Congress Network Advisory Committee (NAC, an umbrella group comprised of library and networking professionals) recently shifted its attention to non-bibliographic databases (which it defines to include full-text, numeric, and graphic data). NAC devoted an entire program meeting to this topic last year. Entitled "Beyond Bibliographic Data," its goals were to gain a better understanding of the term "non-bibliographic" in the library network context and to begin to appreciate the range and potential of such electronic information. Among other things, by the end of the meeting, it was agreed that: librarians must be able to cope with the multiplicity of forms of information; a user interface that will enable scholars and the public to access and display bibliographic and non-bibliographic data files must be devised; standardization and information selection issues remain outstanding; and libraries with local systems and how they affect the relationship of libraries to bibliographic utilities is emerging as a problem: At risk is resource sharing as librarians have known it as attempts are made for economic reasons to seek lower cost alternatives to cataloging on the utilities and thereby not adding expensive cataloging records to a large national database for sharing. The complete proceedings of the meeting are available as part of the Library Congress Network Planning Papers series.

EDUCOM has recently accepted a project proposal, "The Library and the Electronic Document Environment Infrastructure," submitted by Mrs. Avram on behalf of the Library of Congress that calls for a full-scale effort coordinated by EDUCOM, in conjunction with the various stakeholder communities (libraries, researchers, information processors, publishers, professional organization, etc.). The proposal concentrates on three areas of activity:

1) to carry out major studies in three core areas of the electronic document environment:
   a) technology and formats
   What information resources are needed and in what electronic format?
   How will the information be organized for retrieval, transmission, and exchange?
   Will the technologies provide solutions to problems of storage, preservation, and presentation of information?
   b) economic issues and choices
   Who pays for and owns information resources in the electronic document environment?
   c) roles and responsibilities for libraries
   How will the library exercise its role as organizer, classifier, and preserver of information and knowledge in the national network?

2) to devise several structural models conducive to the growth of electronic document environment; and

3) to offer programs, seminars, and publications which present findings about the library in the age of electronic research, production, and publishing.

What has emerged from this discussion is the notion that through the application of technology to networking, libraries will become boundless and that users, by accessing networks, will become patrons of "libraries without walls." Right here in this state, a plan has been issued by the New York State Library which details how all libraries in the state — academic, school, public, or other — can become electronic doorways for citizens of New York. An electronic doorway library would make needed information available electronically to users from any part of the state via links to databases and resource sharing programs with computers.

Research libraries are moving ahead on several fronts through various organizations, both singly and collaboratively, in dealing with non-bibliographic data in a network environment. One of the most promising involves ARL, CAUSE (the association for the management of information technology in higher education), and EDUCOM forming the Coalition for Networked Information. The Coalition will consist of a large and influential group of institutions of higher education, not-for-profit organizations, corporate sponsors, and government agencies. It has set for itself an agenda that includes crafting a set of initiatives to deal with the provision of information resources on the National Research and Education Network. The Coalition will focus on issues related to intellectual property rights, standards, licensing, service arrangements, cost recovery fees, and economic models. So far, as of May 4, over sixty research libraries in the U.S. and Canada, including the Library of Congress, have committed to joining the Coalition.

As we move into the 1990s, it is fair to say that the implications for research libraries of networking and the changing network infrastructure are immense. But, as can be seen, this final decade of the century holds great promise to be an exciting and innovative one as well. And while the task before research libraries in servicing non-bibliographic data in the network setting is staggering, some excellent first steps are being taken.

[1] Presented at the IASSIST 90 Conference held in Poughkeepsie, N.Y. May 30 - June 2, 1990.

NARA Job Announcement - Center for Electronic Records

Job Announcement - March 1, 1991

Within the next two weeks or so, the National Archives will announce one or more vacancies for senior archivists to deal with computer records. Starting salary is $37,294 with all fringe benefits of U.S. Government employment. U.S.citizenship required. If any one is interested, or knows anyone who may be interested, please contact the following:

Thomas E. Brown
Chief, Archival Services Branch
Center for Electronic Records
National Archives and Records Administration
Washington, D.C. 20408
(202) 501-5565
TBROWN@DCUNSN.DAS.NET

Alternative Databases: The Research Resource Division For Refugees

by Les Teichroew [1]
Research Associate
Research Resource Division for Refugees
Carleton University Alternative Databases:
The Research Resource Division for Refugees

Part of the Centre for Immigration and Ethnocultural Studies, the Research Resource Division for Refugees (RRDR) is located at Carleton University in Ottawa, Ontario, Canada. It was established in 1985 to serve as an international archive and data collection agency for scholarly, governmental and field information on refugee resettlement and adaptation. In addition RRDR publishes a quarterly newsletter INSCAN (International Settlement Canada), each issue being devoted to a specific topic on refugee resettlement. RRDR also is actively involved in research on resettlement. To date three studies have been completed: The Southeast Asian Refugee Study - A Report on the Three Year Study of the Social and Economic Adaptation of Southeast Asian Refugees to Life in Canada (1981-1983); The Settlement of Ethiopian Refugees in Toronto (1989); and, the Settlement of Salvadoran Refugees in Ottawa and Toronto (1989). The Resource Division also publishes a series of Working Papers in Immigration and Ethno-Cultural Studies addressing critical issues in the area. Examples include: Immigration and Visible Minorities in the Year 2001 - A Projection - by Dr. John Sammuel and The Mosaic a Generation Later - Issues and Trends - by Dr. Frank Vallee.

The intent and form of RRDR's activities is largely a product of one fundamental characteristic of the field of refugee studies: namely, its state of considerable and continual development and flux. The composition and characteristics of groups of refugees and refugee claimants can change quickly, placing new and unforseen demands upon involved governmental, inter-governmental and non-governmental organizations, immigrant and refugee service centres, and sponsors. What these organizations and individuals require in order to provide services in an appropriate and culturally-sensitive manner is information; but this information needs to be produced quickly, and it should also be readily accessible. As it takes up to two years for information to begin circulating through traditional academic venues, so-called "gray zone" publications offer the quickest source of up-to-date information in the field. As a result, the holdings of RRDR are predominantly - but not solely - comprised of "gray zone" publications. Our holdings also include a limited number of central, but specialized academic sources in the field. In addition, each of our documents has been entered into one of three on-line **textual** bibliographic databases in order to further enhance speed and ease of access to the required information.

"Gray zone" publications are produced by a wide range of organizations, including various government departments, inter-governmental organizations, non-governmental organizations, ethnocultural associations, and other interested individuals, including some academics. Their form is also diverse, including newsletters and periodicals, research monographs, field reports, occasional papers, conference proceedings, pamphlets, posters, and non-print forms, such as videos, films and photographic slides.

There is a considerable breadth and diversity in the range of topics which one may find published in the field of refugee studies. However, RRDR has from the outset focused upon issues of resettlement as they relate to third-world refugees. Areas of particular interest include the situation and experiences of refugee women, the health and mental health of resettled refugees, and the cultural background of refugee groups. However, numerous faculty members at Carleton University can be consulted depending upon the area of research interest. We have made a conscious decision not to systematically collect documents relating to issues of human rights and refugee law, and we also limit the information obtained regarding the political conditions in refugee-producing countries which underlie refugee flight. The primary reason for not focusing upon these two areas is economic: collecting documents in these areas would require financial and personnel resources which we do not possess. Fortunately, two years ago the federal government of Canada established an organization whose mandate was to act as a documentation centre for information on issues of human rights and political conditions in countries of origin. The Immigration and Refugee Board (IRB), through its documentation centre (IRBDC) in Ottawa, Canada and five regional offices, [2] has rapidly obtained a rather diverse and comprehensive collection of documents on these issues - and others - which are readily accessible to government officials, academics and other interested parties. To our knowledge they are the only government internationally to have invested the resources to establish such a collection of

documents. The resources in the IRBDC are complementary to our own: indeed, we often exchange information.

While RRDR purchases a portion of its holdings, many are obtained at no cost from governmental, intergovernmental and non-governmental organizations, ethnic community groups and researchers. RRDR also obtains a significant number of documents through exchange for our newsletter. We have explicitly sought to avoid obtaining documents which are already held at the Carleton University library on-campus; fortunately, owing to our focus on gray-market documents there is little over-lap between the library holdings and our own. However, the library holdings provide easy and rapid on-line access to academic sources in the field - as well as some government documents - thereby extending the range of materials available to researchers in our office. In addition, a number of national libraries exist in the region, enabling RRDR researchers to search for (on-line) and obtain documents from a very wide range of sources. And, while we currently do not have any statistical data sets in our holdings, we would welcome the contribution of such data by any researchers, agencies or governments to our organization.

Although the holdings in RRDR are unique - that is, text-based, predominantly "gray-market" publications on the resettlement of third-world refugees - the format of our on-line records share features similar to those found in many other refugee documentation centres worldwide. This is because we exchange information with participating members of the International Refugee Documentation Network (IRDN) as well as other organizations and agencies. The format for our on-line data follows closely the convention established by HURIDOCS. Although initially intended for documents on the issue of human rights, HURIDOCS has been adapted for use in the field of refugee studies, and also to meet our particular local needs. This format is somewhat different than that used in most libraries, owing partly to the different nature of our respective documents. A number of the participating members of the IRDN have adopted the HURIDOCS format, also with minor modifications.

In numerical terms, our holdings currently comprise more than 6,000 items (i.e. publications) which have been entered into a number of on-line searchable (text-based) bibliographic databases. One of the databases focuses on the condition and experiences of women refugees in countries of origin, countries of first asylum and countries of resettlement. This bibliography is in the final stages of editing and will be published shortly. A second database contains items relating to the physical and mental health of refugees. A third database covers more general issues of refugee resettlement and integration, including economic, cultural, linguistic, and civic and social welfare facets, among others. Ideally, each of the sources in the databases will be fully abstracted and keyworded using the International Thesaurus of Refugee Terminology (developed by the International Refugee Documentation Network, 1989). This should enhance the utility of the databases as the entry of standardized keywords will increase the speed and precision of literature searches. However, owing to limited financial and personnel networks only the bibliography on refugee women has been fully abstracted and keyworded.

Our holdings include items published by international and national non-governmental organizations, governmental and inter-governmental bodies, local ethnocultural communities, research institutes, and interested individuals.

Examples of our international periodicals from non-governmental organizations include ICVA News, published by the International Council of Voluntary Agencies, the International Catholic Migration Commission Newsletter, by the International Catholic Migration Commission, and Refugee Participation Network, by the Refugee Studies Programme, Oxford University. Refugees, from the United Nations High Commissioner for Refugees, and Monthly Dispatch, published by the Intergovernmental Committee for Migration are two instances of our inter-governmental periodicals holdings. National periodicals include the Canadian Ethnocultural Council, by the Canadian Ethnocultural Council, Update, by the U.S. Catholic Conference, Migration and Refugee Services, and Ny Fremtid, by the Norwegian Refugee Council. Immigrant Women of PEI, by the Immigrant Women's Group in Prince Edward Island, Canada, is one example of more local periodicals from non-governmental organizations. Our holdings also include a number of newsletters and periodicals from ethnic associations.

Our holdings of reports, field studies, occasional papers and policy analyses are as diverse. Instances of documents from national governments include The Hmong Resettlement Study, published by the U.S. Department of Health and Human Services, Please Listen to What I'm Not Saying: A Report on the Survey of Settlement Experiences of Indochinese Refugees, 1978-1980, by the Australian Government, and A Follow-up of the Conditions of Unaccompanied Minors and Handicapped Persons among the Refugees from Vietnam Resettled in Sweden, by the Swedish National Board of Health and Welfare. Our holdings also include provincial and state documents such as The Training Needs of Settlement Service Workers, by the Ministry of Citizenship and Culture, Government of Ontario, and documents from local governments, like the publication Information for Young Refugee Parents from the Fresno County Department of Social Services. Our holdings also include a number of publications from intergovernmental organiza-

tions such as Working with Refugees in Somalia Towards a Development Perspective: A Technical Co-operation Report, by the International Labour Office, and numerous reports from various United Nations bodies, including Violence Against the Vietnamese Boat Refugees: An Assessment of Needs and Services from the United Nations High Commissioner for Refugees.

Examples of documents from national non-governmental organizations include Uprooted Angolans, by the U.S. Committee for Refugees and Voluntary Repatriation Programmes for African Refugees: A Critical Examination, published by the British Refugee Council. Making it on Their Own: From Refugee Sponsorship to Self-Sufficiency, by the Church World Service, and Helping Refugee Women Help Themselves: YWCA's Response are instances of publications from international non-governmental organizations.

While one could provide many more instances of our publications holdings, the preceding recitation should have provided some indication of the range of documents in RRDR.

One of our mandates is to make this information widely available to persons and organizations working in the area of refugee resettlement; we often receive requests from organizations world-wide for information in the field. The databases make it relatively easy to process these information requests and, following an on-line search, to forward our findings. In cases where the individual or organization would like to acquire specific documents in our holdings we may, with the permission of the authors, reproduce and forward copies of the publications. In events where this is not possible or feasible, we refer requests to the original publisher.

Unfortunately, owing to limited computer and software facilities in RRDR, the databases are not currently accessible directly from terminals outside of the office and off-campus. On-line access to a read-only copy of the databases is planned for the near future, pending the availability of resources.

Requests for information are received and dealt with at RRDR through a number of routes. Many information requests are forwarded to us by phone, fax and electronic-mail.[3] The simplest of these requests can often be answered over the phone. More complex information requests, or those requiring either a bibliographic listing of publications or the publications themselves, involve sending the information to the requesters by fax, electronic-mail, or through the postal service. A number of requesters also conduct research on RRDR premises. Users of RRDR facilities include students, academic and other researchers, refugee and immigrant service organizations, government departments, and assorted other non-governmental organizations.

[1] Presented at the IASSIST 90 Conference held in Poughkeepsie, N.Y. May 30 - June 2, 1990.

[2] These regional offices are in Vancouver, Calgary, Winnipeg, Toronto and Montreal.

[3] Our fax number is (613)788-3676. Inquiries can also be sent by electronic mail to "neuwirth@carleton.ca" and "rrdr@carleton.ca" .

Data Infrastructure For The Social Sciences In The Netherlands

by Gerton Heyne [1]
IVA, Institute for socila research, University of Tilburg, The Netherlands

Introduction

Computers have become commonplace during the past few years. There has been an obvious increase in, on the one hand, the computerization of data files, and on the other, the analysis and processing of machine-readable data files by computer. The pressing need, particularly by the state, for reliable and up to date policy information and the desire to raise the level of efficiency have been factors in this development. Data stored in computerized form offers greater possibilities for analysis. As a result, scientific research is also able to tap into a growing reservoir of available machine-readable files.

Technical innovations have moreover made it possible for researchers to collect and process research data more quickly and efficiently in order to obtain suitable material for analysis. In part because of innovations in research methods and techniques, researchers are able to carry out complex data analysis not possible before. The technical infrastructure is growing steadily because researchers are increasingly utilizing PCs and mainframes to set up and conduct research and to analyze machine-readable data files. In addition, they also have access to a growing number of networks which can be used to exchange data: for example, the various local university networks and the national SURFnet.

Until a few years ago, when researchers wished to conduct data analysis one of their most important sources was data published in written form. The most important data files were compiled for the state by the Central Bureau for Statistics (CBS). The availability of machine-readable data files on the one hand and the development of infrastructural facilities such as PCs and networks on the other are raising the demand for this information in machine-readable form. The possibility of conducting data analysis grows when researchers can access machine-readable files instead of written publications. In addition such files serve increasingly as "background information." Researchers have access to ever greater amounts of data which can serve as important sources of secondary information for designing and conducting research, because access to files already in use can also, in principle, be made simpler and easier.

The importance of a well-designed infrastructure is evident. Researchers must have easy access to statistical material. Research in the social sciences makes it possible to gain systematic and detailed insights into society. Such insights serve two purposes that are, in fact, directly related. Firstly, research based on statistical information is important for both basic scientific research and for more policy-directed research. Insights gained in this way are of scientific value in and of themselves, but they are also important for efficient management and confident policy-making. Secondly, research serves a fundamental, democratic purpose. A balanced process of policy-making in a democratic society requires that all individuals and parties involved possess relevant information. Such a balance can only be strengthened by a vigorously independent research capacity autonomous of the expertise of the state. To achieve this, however, requires sufficient access to statistical information such as that compiled officially by CBS and other institutions.

In recent years a number of reports have put forward the idea that the data infrastructure used by the social sciences in the Netherlands needs to be augmented (De Bie, 1989; De Guchteneire & Timmermans, 1990). The authors conclude that researchers spend a great deal of time collecting and converting data files before being able to start analysis. They observe the following problems:

- the high cost of purchasing files;
- the restrictive conditions under which files are made available;
- the uneven quality of both the files and the accompanying documentation.

The assumption is that these problems prevent researchers from making adequate use of the data present. Delivery of data files is not optimal. The service provided by suppliers in general, and CBS in particular, is insufficiently geared to the growing demand by researchers for data files and statistics in the form of machine-readable numerical material.

The technical innovations described above have transformed desires and wishes on both the supply and demand side of data files. In addition, priorities have shifted. While the transformation process was still underway, the various parties formulated new goals and

interests that did not always turn out to be complementary. Researchers insist that data files be made available at minimal cost and under the least restrictive terms possible. The policy of CBS, the most important producer of data files, is not entirely self-determined but depends partly on political considerations. This supplier is obliged to comply with political agreements concerning the cost of field work and development and with statutory regulations concerning privacy. Part of the criticism that is being heard, therefore, concerns political leaders rather than CBS.

It looks as if technical and social innovations have set a process of change in motion that has caught the parties involved inadequately prepared. Measures have been taken to secure the interests of both producers and consumers of data files as well as of the individuals whose personal information has been collected. However, these measures often seem to seriously impede the work of researchers.

Design of the present study

Data infrastructure is becoming increasingly important to researchers. Improving this infrastructure is extremely important for the quality of social scientific research. The present study considers the current state of affairs with respect to the availability, accessibility and use of statistical data files[2] in the social sciences[3]. Which files are actually being used in research? Where do the bottlenecks occur in use or loan?

The study comprises an inventory of the use of **external** [4] statistical data files in 1989. The files had to meet a number of criteria formulated beforehand.[5] In order to gain as great an insight as possible into the use of external data files, both the **suppliers** and **consumers** of such data files were approached. This also offers insight into the way in which data files are made available for research in the social sciences. This is the first time that the above-mentioned suppositions concerning the functioning of data infrastructure have been subjected to systematic and relatively large-scale testing by those actually involved in this issue in practice: the researchers.

A number of **suppliers** of data files were requested to provide information about files either sold, lent or offered in some other fashion to researchers in 1989. A number of questions concerning the price, the size and the consumers of these files were added.

A number of users of data files received a questionnaire concerning files they had used in 1989. This also included questions concerning the price and size of the files, time of delivery, problems, quality, etc. Given the large number of research institutions, both commercial and non-profit (including faculties, research groups, institutes and individuals), it was almost impossible to ask all of them to participate in the study. The goal was in any case to approach all relevant faculties. For this reason a number of relevant university faculties and research institutes were selected. An attempt was made to assign each research group or institute a contact person in charge of coordination. Only in a small number of cases did this contact person have a full understanding of every aspect of external data file use in his or her research group or institute. The rest of the time, contact persons were requested to distribute the questionnaires to those researchers who might answer questions concerning the use of external data files.

This paper will provide a short description of
a. the most important information as provided by two institutions that supply data files to third parties (CBS and the Steinmetz archive);
b. the most important results of the survey distributed to a number of data file users.

Use according to two suppliers

CBS and the Steinmetz archive, the Dutch data archive for the social sciences, were asked to answer a number of questions concerning the files that they supplied to researchers in 1989.

Central Bureau for Statistics (CBS)

CBS provided an outline of a two-year period, specifically the microfiles and publication files[6] delivered in 1988 and 1989. The reason given was that in this way chance fluctuations would be less likely to misrepresent the information.

The following deliveries took place in 1988 and 1989:
- 91 deliveries of 20 different microfiles to research institutions;
- 130 deliveries of 10 different publication files, of which:
  * 25 went to research institutions;
  * 66 went to businesses;
  * 39 went to the state.

  Two publication files were delivered 105 times.

Two publication files were by far the most popular, being delivered 66 and 45 times respectively. In addition, eight micro-files were made available more than five times in the period concerned; the Socio-Economic Panel was the most frequent (16 times).

CBS additionally supplied aggregated data[7] concerning the prices paid for various files by showing proceeds per file. It was impossible to deduce from this information how much individual deliveries yielded, or rather, what consumers must have paid for them. For this reason an outline of average prices[8] must suffice. Averages should

be interpreted with the necessary degree of caution, but the average amounts can serve as an indication. In 1988 and 1989 for individual deliveries of eleven microfiles average amounts have paid above 10,000 Dutch guilders (approx. $ 6,060). The data shows that most microfiles are quite expensive. The highest (average) amount for one of these microfiles was 58,800 guilders (approx. $ 35,600). Publication files, on the other hand, cost significantly less. Prices of the publication files supplied most often averaged 630 guilders (approx. $ 380) and 340 guilders (approx. $ 205) respectively.

Deliveries yielded 1.75 million guilders (approx. $ 1.05 million) in total; 1.59 million guilders (approx. $ 0.96 million) came from microfile deliveries. Publication files yielded 0.16 million guilders (approx. $ 0.10 million). The amounts mentioned do not include compensation for additional field work or development costs.

The Steinmetz archive

The Steinmetz Archive distinguishes between research files for the social sciences and weekly public opinion research (weekly questionnaires). The following will deal only with the first category. In 1989 the Steinmetz Archive sold[9] or made a file available 286 times. A total of 111 different files were offered; 42 remained after various volumes or waves belonging to the same file were combined.

The table below indicates how often Dutch and foreign files are used in the Netherlands and how often Dutch files are used in foreign countries.[10]

Table 1. Use in and outside the Netherlands of the Steinmetz archive

	use in the Netherlands	use outside the Netherlands	
Dutch files	110	105	215
foreign files	71	—	71
Total	181	105	286

There are 110 Dutch files and 71 foreign files in all used by Dutch researchers. Almost half of the Dutch files were sent outside the Netherlands.

Three distinct groups of foreign users can be distinguished: universities, data archives and research institutes, comprising respectively 81%, 13% and 6% of the foreign consumption. In the Netherlands users are universities and research institutes, which received 87% and 13% of the files respectively.

Use according to the users

In the survey 70 institutions were requested to complete one or more questionnaires concerning the use of external data files in 1989. The response was low.

Table 2. Response

	approached	response	
university research groups	**44**	**18**	**41%**
- social sciences	31	11	35%
- economics/business	13	7	54%
research institutions (non-profit)	**23**	**8**	**35%**
research institutions (commercial)	**3**	**0**	**0%**
total	70	26	37%

Of the 44 institutions that did not return a completed form, ten stated that they did not work with external data files and can thus be categorized as non-users. The remaining institutions failed to respond at all. The controlled response was therefore upwards of 51%. The 26 institutions, research groups and faculties returned 80 questionnaires in all with information concerning 121 files.

Because of the high level of non-response we do not consider the data ultimately received as statistically representative. The data does not provide an exhaustive view of the use of external data files. In our opinion, the results are nevertheless important for gaining insight into the use of files and the issues related to this use, in view of the number of completed questionnaires and their distribution across various institutions.

Intensity of use

The table below provides numerical data concerning the intensity with which the data files supplied were used. A distinction is made between CBS files and files coming from other supplying agencies.

Table 3 Intensity of use

	CBS	Rest	Total
1. A few analyses shortly after receipt	2%	4%	3%
2. Many analyses shortly after receipt; no others after this	10%	11%	10%
3. A few analyses over a long period of time	16%	37%	26%
4. Many analyses over a long period of time	62%	45%	54%
5. Unknown	10%	4%	7%
Total	100%	99%	100%

Intensity refers to use in time and the number of analyses performed on the file. The figures indicate that the files were generally used intensively. Approximately (54+26=) 80% of the files were used over a **long period of time**, and **many** analyses were performed on about (10+54=) 64%. For well over half (54%) of the files, **many** analyses were performed for **long periods of time**. Use of the CBS files is proportionately more intensive, in the sense that in general **many analyses** were performed.

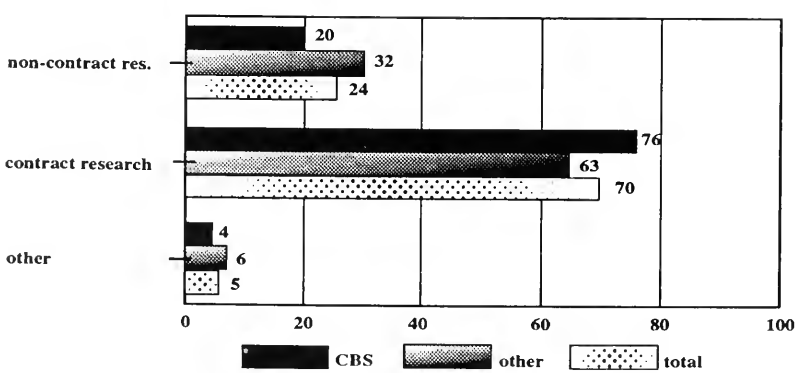
Types of research

The study asked what type of research the files were used for. The respondents were asked how often the file concerned was used for contract and how often for non-contract research. The results showed that files were chiefly used for contract research, in particular the CBS files (see figure 1). Indirectly this was actually a query as to who subsidizes the research (and the files). This information made it possible to indicate which commissioning agencies were directly or indirectly involved in the delivery terms for data files. For non-contract research this is the state. The respondents were asked to indicate who commissioned contract research. The state was named as commissioning agency in 75% of the cases. The conclusion is that the vast majority of research in the social sciences is commissioned and subsidized directly by the state. All in all 33 percent of this was charged to the ministry responsible for scientific research, the Ministry of Education and Science.

Delivery time for files

Questions were posed as to the amount of time that passed between the first formal contact with the supplier and the actual delivery of the file ordered. Delivery of CBS files took a long time in comparison with delivery of the remaining files. Delivery of a CBS file can take well over seven months on the average; delivery of the remaining files averaged little more than a month. Two-thirds of the "remaining" files were delivered within a month; delivery for the rest of these took longer than three months. Only 21% of the CBS files were at their destination within four weeks. Most of the CBS files were delivered within three to twelve months; 21% did not arrive for more than a year.

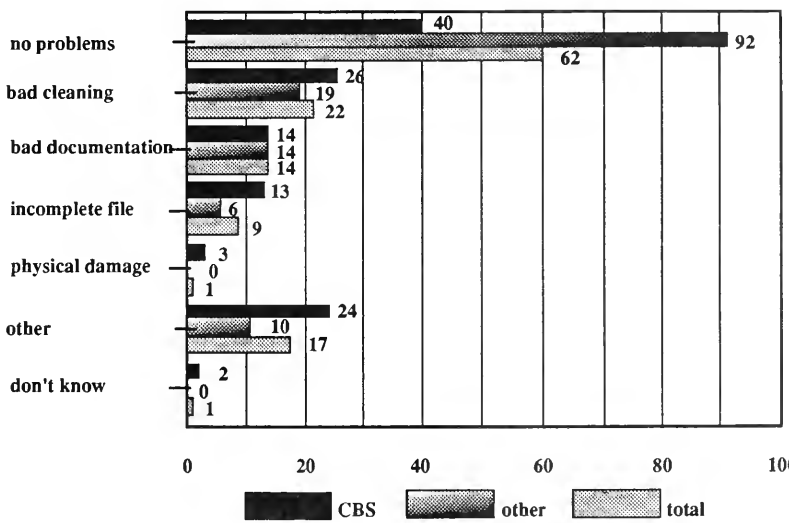
Figure 1: Use of data files



Problems after delivery

Problems involving costs, privacy, etc. come up **before** delivery. The questionnaire included a number of questions about the sorts of problems that came up **after** files were delivered. Of the CBS files, 60% caused researchers problems after delivery (see figure 2). This applied to 40% of the "remaining" files. The CBS files caused more problems for users with respect to cleaning and completeness than the remaining files: 25% as opposed to 19% and 13% as opposed to 5% respectively.

Figure 2: Problems with files after delivery



Quantifying general problems

At the end of the questionnaire the respondents were asked to give their opinion on six different statements describing the same number of problems. This was an attempt to discover to what degree the respondents found specific hindrances problematic. The results are given in figure 3. The data led to the following observations:

1. Almost all scores are negative. This means that all of the aspects illustrated in the statements were seen as more or less problematic.
2. The most important problems concern the **availability** and **delivery** of the files; files are **too expensive, considerations of privacy** hinder or prevent delivery, and when **delivery** does take place, it **takes too long**.
3. CBS file users and those who used other files had widely diverging scores with respect to these important problems in particular. Those who use CBS files find almost every aspect more problematic than those who use other files.
4. On the other hand, CBS scores relatively high on **familiarity**.

Archiving

A data infrastructure that functions well presupposes facilities not only for efficient delivery of files but also for file storage and management. The study asked what happens to files after use. Remarkably, only 80% of the files were archived in any way (see figure 4). The fact that 8% of the files are returned to the supplier and 6% are destroyed may be related to the terms under which the files were delivered; a supplier can, for example, require that files be destroyed or returned after a project has been completed.

Nothing was done with 4% of the files. As this study focuses on files delivered by others, one may assume that the original files remained in the possession of the supplier. Nonetheless, if the same practice applies for files put together by the institutions themselves, then there is an irrevocable loss of material. However, the study gives no definitive answer as to whether this

actually does occur. Only 1% of the work of archiving is contracted out to others, and this although a national archive exists for storing all files that in theory may prove relevant in future social scientific research. Once again, this study concerns files whose original versions are in all probability still with the supplier; copies therefore did not necessarily have to be archived by others. However, by the same token it was by no means strictly necessary to archive files locally, which is what happened in 80% of the cases anyway. For this reason the balance between "archiving" (80%) and "archiving by others" (1%) remains remarkable.

The next question concerned the way in which archiving, management and registration take place. No uniform archiving method appears to exist. Archiving and management take place at different levels; sometimes the individual researcher does it; then again it may be left to someone within the research group, the faculty or the computer center. The lack of guidelines or agreements for storing data systematically hinders access to files that might be suitable for secondary analyses in the future. There is no clear survey of the various locations where files are available.

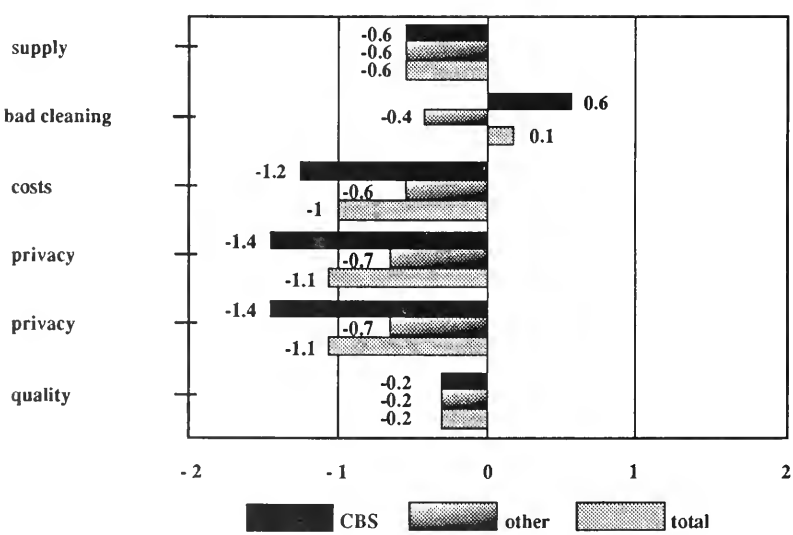
Conclusions

We must first mention that this study focuses largely on the use of external data files and the problems researchers encounter in this use. The most important points of this issue are taken up in this article. Suppliers of data files, on the other hand, are not given an opportunity in this study to express their grievances concerning file users. Although this article says nothing about their complaints, this by no means suggests that they do not have problems with users' unclear, unreasonable or badly formulated wishes.

The study has led us to formulate the following conclusions.

1. Poor response is one of the most important reasons why this study did not result in an exhaustive inventory of the use of data files in the social sciences. It is remarkable that researchers who depend so much

Figure 3: Quantification of problems



The respondents have given their opinion on a number of statements included in the questionnaire. Scoring ranges from 1 = "Agree completely" 3 = "Indifferent" to 5 = "Disagree completely". For clarity the results have been converted into a figure whose minimum score is equal to "problematic" and whose maximum score of 2 is equal to "not problematic".

on the cooperation of others in their work could respond so poorly.

2. The various institutions and research groups within the social sciences have not kept systematic records concerning the use of external data files. Perception of how files are used is diffuse. In this respect, there is a lack of order in the way social scientific research is organized and conducted. In no sense does it fit the image of a professional, well-oiled machine. The data highways see very little orderly traffic.

3. CBS is actually the most constant factor on the data file market of supply and demand. As the most important supplier of data files, CBS has, firstly, a wide range of files to offer and secondly a number of clearly stated delivery terms. This gives CBS a well-defined policy concerning the delivery of data files, making it the cornerstone of data facilities. The demand side, or rather the research world, has little to offer in return. As mentioned before, it is a diffuse and unorganized field in which mutual interests and viewpoints are difficult to formulate. This unequal situation hinders coordination between the parties concerning delivery terms for the files.

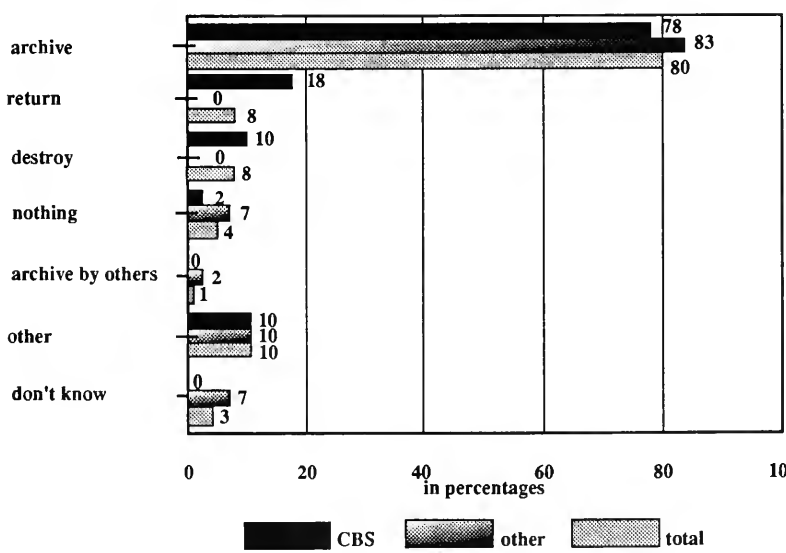
4. The information provided by CBS makes clear that a demand with great purchasing power exists for a select number of files in the social sciences. CBS's files are better known than files offered by other suppliers. The study also confirms what has been observed in the literature: that researchers find the limited availability of microfiles due to cost, considerations of privacy and delayed delivery problematic. These problems are particularly in evidence with respect to CBS files.

5. The demand for Steinmetz archive files is quite specific. In 67% of the cases, the Steinmetz archive delivers files containing voting research data. Remarkably, almost half of the Dutch files go to foreign countries, and an important portion of the files used by Dutch researchers are from abroad.

6. The survey reveals that 70% of the research is on a contract basis, and that almost all of this is financed by the state. Given that the state also finances the remaining research either directly or indirectly, we can conclude that the state subsidizes almost all social scientific research for which data files are purchased.

7. The fact that most of the data files are used in contract research, together with the observation that non-contract research is much less likely to use CBS

Figure 4: Archiving



data than data from other files, indicates the difficulty, if not impossibility, of acquiring costly files given the limited financial resources of the various research groups.

8. Because most of the files are ordered for contract research, it may seem acceptable to pass the costs of the files on to the users. However, it is also important to note that while certain costs may be included in the price of the file, other costs should not be.

9. Local storage and management of data files is conducted at various levels. There are no guidelines or agreements concerning systematic data storage. This hinders access to files that might theoretically be suitable for secondary analysis.

RECOMMENDATIONS

Technical innovations have made the processing of machine-readable statistical files increasingly important for social scientific research, a development requiring reflection upon the supply of files and the use of new technical facilities.

1. This reflection can take shape in the form of coordination between the suppliers and consumers of data files. Segers (1990) has already observed that the research world is too unorganized to set up a structural dialogue with suppliers; it must first gather its forces and negotiate with suppliers over new delivery terms and a broader range of user options. Segers recommended establishing an independent organization to represent the research world. As part of its task this institution would coordinate the desires and wishes of the users on one side and CBS on the other. On the basis of this coordination a number of standard agreements could be formulated.

2. Clearly CBS must play a vital role in improving the data infrastructure. Its cooperation is necessary in implementing a number of possible adjustments concerning:

- first of all, the **delivery policy** of CBS. Too much time is lost bargaining over terms of delivery. Delivery of microfiles is often laborious, partly due to the disclosure risk; as a result, delivery is expensive and irregular. Prices increase because files have to be custom-made every time. Both researchers and CBS would be better off if they switched to making standard deliveries of the most important data, perhaps in the interim stage in the form of publication files. These would be "stripped" microfiles.

- second, the **cost of microfiles**. Besides the risk of disclosure, this is the most important barrier to delivery of files. The standard delivery of data mentioned above, in the form of inexpensive publication files, may help to solve this problem, but the cost of microfiles must be reconsidered further. The main idea would be to generate a buyers' demand within non-contract research; in other words, to take a number of measures that will facilitate the acquisition of files by university research groups. The recommendation made by De Guchteneire & Timmermans to establish a central budget for the purchase of data files might serve as a point of departure.

- third, the **use of other media** than what has been used to date. This involves creating facilities destined for long-term use, for example CD-ROM. This initiative clearly supports the standard delivery of important data. CBS has assumed that for the time being such media are more suitable for the dissemination of aggregated data than for the distribution of individual level data, for example all statistical data concerning municipalities. Mention must be made here that university libraries will no doubt be unavoidably drawn into this issue. When libraries purchase and make such information available, it becomes available to a much wider audience of users.

3. Data infrastructure would profit by improved "signposting". Users or potential users must have better information about which files are to be found where, and how and under what conditions they can be acquired. At the moment most researchers do not know which paths to follow to get this kind of information; often they may not even know that such paths exist. Consequently, policy should aim at improving information on user options; in other words, documentation, etc. with respect to finding and acquiring the files.

4. Researchers should have a service organization available to them for information and assistance in finding and acquiring data files. In the United States in particular an increasing number of universities are beginning to set up and develop **local** data libraries. In most cases these libraries still form part of a "regular" library. Such libraries perform the following tasks:

* systematically storing used data files;

* entering bibliographical information on these files into a computerized catalogue;

* providing services to researchers or students, including:

  - searching for particular files
  - using the files
  - analyzing the files.

Besides offering advice, other administrative duties can be included as part of the service:

- performing the tasks mentioned above,

- finding and retrieving data files present elsewhere [11] upon request.

5. De Guchteneire & Timmermans proposed establishing a service organization which would perform the following tasks:

* gathering and developing expertise directed at making sensitive material anonymous[12]

* developing techniques through which sensitive material can still be made available (for example, by working with test files),

* advising researchers.

To these tasks may be added:

* designing sound rules concerning liability and other organizational terms concerning the safety of the files.

As in point 4, the issue here is service to researchers. We advise describing the possible functions and tasks of a service organization in a follow-up project, and exploring which of these tasks should or could be performed locally or nationally.

6. To summarize, in shaping a new or adapted data infra-structure attention must be given to the following points:

* organization and representation from the field;

* conditions for making data files available;

* the way in which the files are made available;

* guidelines for data file storage and management;

* service for researchers and students.

An important consideration in describing the infrastructure is the current and future position of the remaining supply agencies, in particular the Steinmetz archive.

Inconclusion

Technical innovations have made it increasingly possible to produce computerized data files, make them available and analyze them. Data can be analyzed more often and from various perspectives. For this reason, we recommend that CBS files be made available without many restrictions to as broad a sector of researchers as possible, and that different types of analysis be conducted by various institutions. Because CBS produces statistics, it must process the collected data in a number of ways. There are so many different manipulations that can be performed on files that it would be impracticable for CBS to carry out them all. This is not to say that CBS would no longer have to perform manipulations, on the contrary. However, microfiles should be made more readily available to a larger group of researchers. Wider delivery of CBS files would serve not only the interests of researchers, but of society at large.

Adequate construction and efficient use of the data infrastructure requires the involvement of builders, managers and users. Technology on the one hand and sound agreements on the other must lead to the development of an infrastructure that allows researchers to make high-speed use of the available "data highways." At present, however, unexpected roadblocks and no trespassing signs have been thrown up. Regulations concerning availability must be amended in order to give data files the right of way. The interests of all parties involved must be kept in mind. It is time to take measures and come to agreements: a well functioning infrastructure is a necessity in a modern and democratic society.

The respondents have given their opinion on a number of statements included in the questionnaire. Scoring ranges from 1= "Agree completely," 3= "Indifferent" to 5= "Disagree completely." For clarity the results have been converted into a figure whose minimum score of -2 is equal to "problematic" and whose maximum score of 2 is equal to "not problematic."

Bibliography

Bie, S.E. de, Wetenschap en statistiek, standpunt van de vereniging van Onderzoek Instituten inzake de levering van CBS microdata ten behoeve van sociaal wetenschappelijk onderzoek. Leiden, VOI, 1988.

Guchteneire, P.F.A. and J.G.M. Timmermans, Wetenschappelijk gebruik van overheidsdatabestanden. Amsterdam, Swidoc, 1990.

Segers, J., De data-infrastruktuur in Nederland: snelweg of zandpad? Presented during the "Statistics Day" conference, 1990.

Steinmetz archive, Data catalogue & guide. Amsterdam, 1990.

[1] Presented at the IASSIST '90 Conference held in Poughkeepsie, NY, United States, May 30-June 2, 1990.

[2] These files were set up for statistical analysis and can theoretically be used for planning and policy. As opposed to data in an administrative file, data stored in a statistical file is not altered again except for aggregations and scale or index constructions. For example, data included in the "Population statistics" file undergoes no further change once they have been collected, whereas data in an administrative file such as the register of births, marriages and deaths is altered whenever a resident whose personal information is included in the file moves or receives a new passport or driver's license. In statistical files, just as in administrative files, informa-

tion is often collected at the level of the individual. In the analysis of statistical files, however, the object is manipulations at the aggregate or case level. Data from administrative files, in contrast, is used at the individual level (De Guchteneire & Timmermans, 1990), as the above example makes clear.

[3] The **social sciences** encompass a wide range of fields, and the borders are not always sharply drawn. Even the Dutch term "gamma sciences" does not solve the border issue. In the context of this study the following categories were used:

- Economics (microeconomics, macroeconomics, business economics and econometry) and business administration;
- sociology;
- psychology;
- education;
- political science;
- public and policy administration;
- environmental planning, geography and related studies.

[4] These are files whose data was **not** collected at the initiative of the institution itself, but were acquired or purchased from others. Not included are files whose data

- was collected by the institution's own field workers;
- was collected by others at the request of the institution.

[5] The files had to meet the following criteria:

* the data was in a file; the study specifically did not focus on tables that were delivered;
* the files were stored in such a way (tape, diskette, CD-ROM) that they could be accessed by computer (that is, as a datamatrix);
* the files were in the possession of the researcher, research group or institution; on-line external files were not included in the study;
* microfiles (data at the level of persons or households) had to include at least 1000 research units; mesofiles or macrofiles (data at the level of companies, institutions, sectors or countries) needed at least 100 research units;
* the files were used for research in **1989**;
* the files were used for research in the social sciences.

[6] In **publication files** the research data, anonymous or not, is reworked in such a way that disclosure of individuals is almost entirely ruled out. **Microfiles** are also anonymous data files at the individual (or household) level, but they include such detailed information on variables that by making intelligent combinations, disclosure may be possible. The ethical code maintained by researchers does not permit such disclosures. Microfiles are only supplied under the terms of an agreement; publication files are offered on the open market. Unlike publication files, microfiles are supplied neither to government agencies serving a public administrative function, nor to business, but only to research institutions, meaning universities and the Planning Bureaus.

[7] For example, the response states that ten deliveries of a particular file yielded 375,000 guilders. It is not clear whether this is two large deliveries yielding 300,000 guilders and eight others yielding 75,000 in all, or another combination.

[8] CBS was unable to meet our request for a survey of purchase prices of the most complete files. Because negotiations are underway concerning policy adjustments with respect to data file availability, CBS found it an inopportune moment to provide price information which would become outdated in the near future.

[9] The weekly questionnaires were offered 226 times in all. These were 226 different questionnaires.

[10] These are files delivered by the Steinmetz archive.

[11] Various developments in the United States have taken place in close collaboration with the International Consortium for Political and Social Research (ICPSR).

[12] CBS comments here that "anonymous" is hardly a viable term. It is not only a matter of sensitive data files but also of data that cannot be identified at the individual level, and of non-disclosure, according the CBS. The Bureau states that a high level of expertise has been focused on this issue and that the Bureau, affiliated institutions outside the Netherlands and certain foreign research institutions are continuously developing greater expertise.

"Archival soundbites, footage, and photographs — past, present and future: The perspective of a documentary filmmaker and sociologist."

by Karl Schonborn Ph.D.[1]
Professor of Sociology ,California State University
Hayward, California

Introduction

Documentary and educational film-makers trying to express sociological concepts have always had a great appetite for archival footage and photographs. In recent decades, this has come to include numbers and statistics and soundbites, too. This appetite will increase in a quantum fashion when hypermedia[2] (also called multi-media) takes hold as an educational, instructional tool.

I am a sociologist and the writer, producer and director of six major educational documentaries which have been rented and sold to universities and organizations across the United States. As a consequence, I have spent a good deal of time tracking down footage, photographs, and sounds to include in documentaries. Therefore, I would like to share some thoughts about the storage and retrieval of such items - in the past, currently, and in the future.

My remarks will pertain primarily to documentaries and educational films, but they have relevance to many other instructional uses of such information. My favorite definition of the documentary is that it is a file or tape made with no love story, no plot, and no anticipation of profit.

A more serious definition, befitting the archival focus of this paper, is presented by legendary documentarist/critic John Grierson: documentaries entail "the creative treatment of actuality." Relatedly, documentary pioneer Dziga Vertov claimed the documentary's task is to capture "fragments of actuality" and combine them meaningfully.

Radio and television news still use the term "actuality" to refer to the real world sights and sounds they record. Documentaries and other educational films and tapes generally present "actualities" within a frame of reference or provide some interpretation.

Past

Regardless of the kind of documentaries they made, film-makers have almost always spent time chasing down the images, words and sounds they needed to make their point(s). Erik Barnouw talks of the documentarian as biographer (e.g., P.M. Adato and his "Georgia O'Keeffe"), historic chronicler (e.g., Barbara Kopple and her "Harlan County, USA", explorer (e.g., Robert Flahrety and his "Nanook of the North"), promoter (e.g., Frederick Leboyer and "Birth without Violence"), and guerrilla (e.g., Peter Davis and his "The Selling of the Pentagon").

"Biographers" obviously utilized a good deal of archival data, whether more traditional ("Paul Robeson: Tribute to an Artist" - 1980) or innovative ("Wasn't that a Time!" - 1982 - a celebration of the singing group, The Weavers). These works are not to be confused with "docudramas" which likewise use archival data but which are really historical fiction rather than documentary (Barnouw, 1983:309).

And "historic-chroniclers" likewise were heavy users of archival material. Since World War II, film archives have proliferated because many new countries felt film archive collections would underscore their unique cultural traditions and beginnings. Film-makers and collectors often gave their holdings to such collections.

Revisions of historic dogma often resulted as documentarists using these collections sometimes found "it just wasn't so." Revisionist history got a boost from such debunking documentaries as "World at War" (1973), a re-examination of WWII by Thames Television; "Men of Bronze" (1977), a chronicle by William Miles of a WWI unit of black Americans who fought heroically for the French after being refused by General Pershing who only wanted to command white soldiers; and Connie Field's "The Life and Times of Rosie the Riveter" (1980) which chronicled the important contribution American women made to winning WWII (Barnouw, 1983:308).

On a personal note, I remember in the 1970's going to the Picture Collection at the New York Public Library whenever I was "East" — being a Californian — because I needed a picture of Freud's mother or a shot of Babe Ruth's wife for my work. The remarkable thing about this collection besides its breadth was the easy access to it. An out-of-state person such as myself could get a lending card with no trouble and — even more amazing — could borrow numerous photos (usually mounted on pictureboard) for a long period of time, taking them 3000 miles away and mailing them back as I usually did.

In the past, before computers, keeping track of all the footage, still photo inserts, sound effects, etc., for a documentary was a monumental task. Just coordinating the search for such materials was difficult since one often had to retrace ground all over if the image proved impossible to duplicate or the sound was filled with "noise" or other impedimenta.

Present

The advent of the computer meant paper and pencil lists and scores of production assistants could be dispensed with. Herewith is an accounting of how storage and retrieval works in many contemporary documentary projects. Since the film making process is fairly well known (strips of film spliced together, editing tables, A & B rolling, etc.), the focus here will be on the video-making process.

Computer list management and time-coding have made the handling of large numbers of actualities, words, numbers, images and sounds much more manageable for documentarians and makers of instructional videos.

Documentary-makers store and retrieve tape footage of actualities, etc. by means of "time-coding".[3] Using a time-code generator, they put down an electronic signal along the length of the tape(s) on which they have recorded their needed information/data. This creates a unique "address" (in hours, minutes, seconds, and frames) for every bit of material on the tape(s). Time-coding is often done when copies of the original tape are made. (Generally, people work with copies while composing their work prints or "rough drafts" rather than risk erasing or damaging their original tapes. The originals are used again when the final print or "edit master" is created.

When time-coded tape is played, the "address" is seen in a window — usually at the top of the TV screen. This 8-digit address lets the documentarian manually or electronically find the exact place on the tape he or she is looking for. Time-coding is a frame accurate version — with electronic markers — of the counter gadget seen on less sophisticated VCRs.

Lists of the "addresses" (locations) of the start and end of all tape segments can be put into a computer which then allows documentarians to pick and choose the footage, images, sounds, etc. he or she desires. A list of the desired segments can then be used to tell editors (or sometimes automated edit machines) how to assemble the documentary.

The advantage of computers in present-day documentary-making is that they greatly speed up the process of locating and retrieving moving images - as well as still pictures, sounds, graphics and music. They also facilitate the ebb and flow of decision-making re.***sic*** what to include and exclude from the final cut of a documentary — heretofore, one of the most onerous tasks confronting the producer.

Some well-financed documentarians utilize videodiscs,[4] especially for storage and retrieval of still images. They shoot pictures using a single frame 16mm film camera and then transfer the images to videotape with a telecine (essentially a film and video camera combination). Most of them have then had to send the tape[5] to a special lab to make the laser videodisc. This could take days or even weeks.

Mention should be made at this point of several sources of image data for contemporary documentarians. Slide houses and stock footage libraries carry images of all sorts which can be bought. Unfortunately they are fairly expensive sources. A major newsmagazine also sells its hard-won still pictures on the open market and some news stations likewise sell footage. These latter sources — while also expensive — hark back to the "morgue" tradition started by newspapers. To be ready for any breaking story, and especially the deaths of famous people, newspapers created "morgues," i.e., files of photos and newsclippings of selected people and organizations.

Finally, sound effects are sold by the "needle drop" (a holdover from phono records); and, of course, canned music can be bought for a song and copyrighted music for an arm and a leg.

Future

The future is already with us in some ways given that "digitization" of audio and video are current technologies. The main problem for documentarians is that digitized video is currently very expensive.

In the high resolution, high fidelity age we live in, the shortcomings of analog [5] stored and retrieved video images makes digitization attractive as the technology of the future. (Most analog video becomes fuzzy after several generations — copies of copies — but new "digitized" video continues to have sharp video resolution after scores of generations.

Digitization

To digitize an image, it must be scanned using an electronic camera and a computer. Each detail on the image — including light and dark variations — is assigned a number that is stored. If high resolution is desired, then digitization gets very expensive because millions of numbers must be stored. (A 5¼" disc can only hold 2 or 3 images.) And color images require extra

memory, often a megabyte at the bare minimum.

Today, it is more cost-effective to use analog information. One can store 54,000 video pictures on a laser videodisc. But digital compression techniques and other technological breakthroughs should probably make digitization the technology of the future.

Digitized images can come from almost any source, but it helps to start with high resolution pictures. A character generator is used to put an identification "address" on each picture. This is entered on the original image as well as onto a computer using a list management program. Documentarians are then able to enter, find, and extract data from the database.

When recording still frames of video on a disc, care must be taken that there is no dropout, jitter, or chroma crawl. Video encoders and filtering processes can sometimes deal with these problems.

Once an image has been digitized, one can play with it — erase parts of it, increase it, alter it any way ones wants. With high speed digitizers, it is possible to change things in a moving video scene. (High-speed digitizers — called frame-grabbers — can capture images in one-thirtieth of a second. Conventional digitizers may take several minutes, especially if quality reproductions are desired.)

This ability to partially erase or increase a digitized image will solve one problem documentarians face: encountering an image that is so cluttered with distracting elements that the key element is missed or one that is too small to be clearly shot. Even a micro lens cannot always capture the essential element. For example, in my "Stigma" documentary, being able to zoom in on a keloid scar (a thick fibrous scar) on a person would have eliminated my having to consult a medical journal for an extreme close-up picture of a keloid scar.

Organization

The continuing explosion of audio and video information and "actualities" means that archivists and librarians will have to have highly organized systems for selection, storage, and retrieval of audio and visual information. Like a family trying to find a shot of Great-Aunt Betsy in a shoebox of color slides or in a cabinet of home movies, they might easily become overwhelmed. All this assumes that archivists will have the proper training or help to guarantee the "quality" of the images and sounds they store. (Resolution, clarity, audio and video levels, lighting, composition, and security are part of the "quality" issue.)

Let us look closely at the matter of "selection," especially the challenge of selecting which "actualities" to store. While we might debate whether the number of truly "significant" events and verbal utterances has increased in recent times, there is no question that the record of events and words has. Professional photographers, journalists, media departments, and — yes, amateurs — record countless scheduled and unscheduled events, speeches, and goings on. Most amazingly, amateur recordings, such as the Zapruder film of the Kennedy assassination which was rare and freakish in 1963, are now common. Tourists and everyday people with AV gear record planes crashing, bridges collapsing and people being shot.

There are important questions to answer: Who will decide from the avalanche of possibilities which events, speeches, etc., are significant and therefore worth saving electronically? Probably committees composed of historians, social scientists, journalists and the like.

And what criteria will they use? Will archivists automatically save the inaugural and farewell speeches of anyone who has attained a certain political level, say senator or governor? (This would probably be too status and stratification bound and would mean non-establishment speeches (like Martin Luther King, Jr.'s "I have a Dream" speech) would not be stored.)

Will certain people and concepts be favored during certain periods because of their cultural centrality or importance? Maybe early TV shows and personalities ("Mickey Mouse Club," Phil Silvers,) would be emphasized for the 1950s, counterculture happenings for the 1960s, professional athletes for the 1970s, and entrepreneur and CEOs for the 1980s. No matter what criteria are used, the selection procedure is bound to be difficult.

Deliberations over what is significant or not quickly get into issues of "reality." It seems as if more Americans are able to identify certain TV commercials than identify the state of Ohio on a map. (Ditto for certain Hollywood movies.) Apropos of this, there are Clio Awards (Clio was the muse of history) given each year for outstanding ads of various sorts. As a consequence, archivists will have no trouble documenting our "reel" history (according to Madison Avenue and Hollywood). It's just our "real" history that will be difficult to document.

Some further thoughts and recommendations for future information storage and retrieval, though from the limited perspective of a documentarian (I base these recommendations partly on the assumption that everyone will someday be assembling words, pictures and sounds — for institutional use or for personal use (much as people assemble scrapbooks and photo albums today):

- Audio and video data should be stored digitally on videodisc. Film and videotape are too fragile, necessitating recopying every so often because of oxidation and other chemical processes.

- Audio and video information might best be cataloged and stored by social science discipline. These databases would resemble the CD-ROM [6] databases which presently exist for some of the social sciences: e.g., PsycLIT (the APA's Psychological Abstracts), ERIC (Educational Resources Information Center), and Infotrac.

- Mechanisms similar to those used by the United Nations to gather and report data might be used to insure that important audio and video data around the world are preserved and made accessible. The database might be organized along the lines of MEDLINE (published by the National Library of Medicine) which is international in scope. If possible, the databases should not overemphasize Western and European audio and video.

- All databases might be updated quarterly and then reassessed at the end of each year and each decade — much the same way that newsmagazines summarize the year and decade in words, pictures, and the like.

Conclusion

Over the years documentarians and others making educational films have utilized increasingly sophisticated means to obtain the archival footage, photographs and sounds they needed for their work. The amount of this archival data has grown exponentially, and the need to access it will too as interactive learning and hypermedia find their own niche alongside documentaries.

Digital storage on videodisc probably represents the most realistic way for these immense amounts of data to be stored and accessed easily.

Much ongoing effort will be required to select, record, and organize the untold numbers of words, numbers, images and such. The cataloging, tracking, and safe-keeping of audio and video data, though should be well worth it. Nothing comes quite alive like the voices, gestures, and actions of the greats (and even the not-so-greats) of yesteryear. In a sense, archivists of the future will be in the business of freezing people and then bringing them back to life - but electronically rather than cryonically.

Bibliography

Barnouw, Erik, Documentary History of the Non-Fiction Film, Oxford University Press, Oxford, 1983.

Hardy, F., Grierson on Documentary, Praeger, Praeger, 1971.

St. Lawrence, J., "Still the Image," Videography, February 1990.

Stevens, George, "Applying Hypermedia for Performance Improvement," Performance and Instruction, July 1989.

[1] Presented at the IASSIST 90 Conference held in Poughkeepsie, N.Y. May 30 - June 2, 1990.Professor Karl Schonborn, California State University, Hayward, California 94542, 415 881-3173

[2] Hypermedia (or Multimedia) refers to the coordinated use of more than one medium - sound, video, animation, graphics and text for example. Moreover, hypermedia can mean all these media plus "interactivity," where the medium responds to the user and vice versa. Users, typically, explore hypermedia materials at their own pace, moving in different directions through the information, creating their own interpretations, involvements, experiences.

[3] Time-code refers to the 8-digit address code used to identify each video tape frame by hour, minute, second and frame number (allows frame-accurate precise editing).

[4] Videodiscs are like 33 1/3 phonograph records and they store video information. Often used for instant "replays" in sports broadcasts, videodiscs allow for user friendly instant access to information as well as slow motion, freeze-frame, and interactive video effects. A major advantage of discs over tape, though, is there is no diminution of picture sharpness when dubs (copies) of the image are made.

[5] In analog processing, an electrical signal varies over a continuous range to represent the original audio or video that is being reproduced. In digital processing, electrical impulses are either on or off - just as they are in computers. One digit (0 or 1) is generally represented by the "on" state of an electronic device while the other is represented by the "off" state. One can store visual images on laser-based read/write optical discs as analog information or digital information, i.e., as video or as computer data.

[6] A CD-ROM (Compact Disc-Read Only Memory) allows high density storage, having a capacity of 550 Megabytes (roughly equivalent to 1,500 floppy discs or 100,00 pages with 5,500 characters per page).

The Henry A. Murray Research Center: Alternative Data Sources - Unique, Yet Less Visible Archives and Programs

by Erin Phelps [1]
The Henry A. Murray Research Center
of Radcliffe College

Background Information on the Murray Research Center

The Henry A. Murray Research Center of Radcliffe College was established in 1976 as a national repository for data from the fields of psychology, psychiatry, sociology, anthropology, economics, political science, and education. Unlike most other data banks, the Murray Center archives original subject records as well as coded, machine-readable data. These original records often include transcripts of in-depth interviews, behavioral observations, responses to projective tests, or other information that can be used profitably for secondary analysis. This makes possible a restructuring of the subject records and mitigates the degree to which one is locked in to the theoretical assumptions under which the data were collected.

In spite of the clear advantages of making raw data available, most data banks offer only coded computer data. Some major longitudinal studies, for example the New York Longitudinal Study (Thomas and Chess, 1977), are archived in a manner that allows access to the records by outside investigators, but in general each of these studies is housed separately. To our knowledge, the Murray Research Center is the only repository that is designed to offer a wide range of longitudinal data sets, many of which include raw data.

Data sets may be reanalyzed in order to explore questions other than those addressed by the original investigators. Reanalysis may involve recoding of the raw data. Data from studies that employ comparable designs and instruments may be combined to provide a larger, more varied sample than would be possible otherwise. In addition to reanalysis, these studies can be used as baseline data for replication studies in order to assess the effects of social change. In many cases, samples can be recontacted for further longitudinal follow-up. The availability of samples which can be followed up encourages the collection of longitudinal data that would otherwise be too costly or difficult to obtain. Furthermore, the data collected early in this century can be used to address questions of interest to social historians. Reviews of existing data can facilitate exploratory research by allowing a researcher to refine research questions, to assess the best means for addressing those questions, and to develop new research instruments or coding schemes. In addition to using the Murray Center's data sets in the ways mentioned above, faculty and students use these data in courses to illustrate how the data are collected and analyzed.

The resources of the Murray Center are open to undergraduate and graduate students, faculty, and other researchers from around the world. Staff members conduct several introductory workshops each semester and are available weekdays to provide individual consultation on using the Center's resources. Users are not charged for access to the data. They are charged only for special services they require (such as duplicating) and for the cost of computer time. Our Guide to the Data Resources of the Henry A. Murray Research Center is available for $12 and provides information about the more than 190 data sets that are completely processed and those that are still in the process of acquisition. In addition, an Index to the Guide is available for $5, which includes a detailed listing of the methods of data collection, content areas, and an index of the data sets according to these characteristics.

An important advantage of the Murray Center is that the archive is located within an active research center. The Murray Research Center offers staff assistance to data users, seminars and conferences on methods for making productive use of existing data, and a visiting scholar's program which hosts several researchers each year in residence at the Center. This kind of setting fosters intellectual collaboration and helps to ensure that the Center's data resources are used, and used well, for important new research.

Data Formats

Original Records

We currently provide information in two forms: 1) copies of original paper records and 2) coded computer readable data. In addition, we are beginning the process of microfilming paper records, and where possible, converting mainframe computer files to desktop computer floppy disk format. We are also investigating the possibility of archiving videotapes of social interaction.

In the case of studies that include original subject records, processing the data includes several steps. The first is arranging for duplicating and shipping of the

records held by the original investigator, checking the quality of the photocopies and obtaining replacements when necessary, and removing names and other identifiers. Then the material is organized by subject and measure and stored in acid free folders and boxes. A complete inventory of the material is created in order to document the occurrence of missing data, and a detailed documentation is written, describing the data collection procedures and other information about the data set that a user might need. Finally, a "user file" is created that contains the legal agreement between the contributor and Radcliffe College, the inventory and documentation, copies of measures, publications, and information on machine-readable data.

With funding from the National Institutes of Mental Health and the MacArthur Foundation, we have begun to create microfiche copies of original record data. The availability of microfiche copies will make the data more easily transferrable outside our local area. For reasons of confidentiality, however, use of raw data off-site is not allowed without the written consent of the data contributor and will be allowed only when the data are completely anonymous with respect to the subjects' identities. This rule is designed to protect the privacy of the study participants. In some cases, however, when the subjects' identities could not possibly be revealed by the records, data contributors do give permission for off-site use. In such cases, the Murray Center's clerical staff must photocopy all of the material to be sent, since using a copy service might result in loss of the records. This is obviously a very expensive and cumbersome procedure. With microfilm records available, the process of screening and controlling off-site use would be conducted with equal care, but in cases where no risk to confidentiality is involved, the ease of transfer would be greatly enhanced.

Computer Data

In addition to original subject records, most of the data sets include tapes of computer-accessible data. Data tapes are checked against codebooks, and the necessary data cleaning is completed either by the contributor or Center staff. Our current method of processing a computer data file depends upon the form in which the data are received. If a SAS file or an SPSS system file was created on a mainframe that is compatible with our mainframe, processing simply begins with the given file as it is. Commonly, however, we write a computer program in SAS or SPSSX to read a raw number file. In either case, data are carefully checked. Two forms of computer output are generated and checked against any original interview or questionnaire data accompanying the tape. The first is a frequency distribution of all variables. A computer codebook and any other documentation is used to check for out-of-range values, correct missing value designations, and appropriate labeling. Where original data are available, a case-by-case listing of actual values for each variable is generated for a random sample of cases. This listing is compared with actual questionnaire and/or interview data to check for consistency between the two forms of data.

Computer tapes of machine-readable data are routinely copied and mailed to investigators wishing to use them. Users are only charged for the cost of the tape and the transfer. Computer data are already used predominantly off-site and are made available as SAS files, SPSSx system files, portable files on computer tapes, or sent via communications networks such as BITNET or ARPANET. During the past year, requests for data in floppy disk format have been received, a trend that seems likely to continue and increase as desktop computers and appropriate statistical software become more widely available. Because of this, we are beginning to transfer the machine-readable data from many of the Murray Center's data sets to floppy disk format for use by desktop computers. This should facilitate use of the Center's data, as well as decreasing the cost of both transfer and analysis.

Videotape Data

Through a recent grant from the MacArthur Foundation, we are exploring the feasibility of adding another dimension to our archive — videotaped data. Videotapes of social interactions most often record not only those behaviors that are pertinent to the original investigation, but also record a wide variety of additional behaviors which may be coded by researchers investigating other aspects of human interaction. In addition, new systems for coding interaction data are continually being developed, and all require coding of pilot video data to establish preliminary reliability and validity. Existing video data can often serve as pilot data for testing a new coding system.

Because it is extremely time consuming and costly to collect observational interaction data and the data cannot be analyzed exhaustively by one investigator, these data are currently being shared among researchers. The private transfer of data from one place to another is unregulated so generally there are no clear guidelines for who may view the tapes and what conditions viewers must adhere to. An additional problem with private sharing of these data is that many potential users are not aware of the availability of the data sets that may be appropriate to their needs. Thus, in many ways video data seem ideal for inclusion in a data archive such as the Murray Center, which emphasizes the study of lives in context through reanalysis of original records.

Individual researchers and human subjects review boards are just beginning to recognize the ethical problems involved in sharing data that cannot be completely deidentified. A variety of important ethical issues must

be considered regarding the confidentiality of the subjects in video studies. For example, should parents be allowed to give consent to the archiving of identifiable video data of their children? Should subjects be required to view their videotape data before allowing the data to be placed in an archive? These ethical issues exist for any participant in data sharing, whether a private researcher or an institutionally affiliated data archive, although the issues are somewhat more complicated in the case of an archive.

This year we are spending some time thinking about whether or not we should begin a major program of archiving videotaped social interaction data. The first step is to clarify the ethical issues by ascertaining the current range of policies at major universities regarding sharing of these data, determining the possibility of creating procedures for safeguarding the confidentiality of subjects, and making recommendations regarding ethical policies for both private sharing and archiving of video data. If we conclude that it is possible to resolve the many ethical problems that exist, we will begin identifying data sets for potential acquisition.

Other Services

Workshops and Conferences

In addition to archiving data, the Murray Research Center sponsors workshops designed to draw attention to its data resources and provide training in the skills needed to carry out effective secondary analyses. Our plan for the next three years is to offer at least three workshops each year in order to facilitate the use of the Murray Center's archive for theoretically innovative research. We have already begun this expanded outreach to the research community. Within the past year four workshops were scheduled in connection with the Murray Center's recently acquired longitudinal studies. They were: 1) A workshop on methods for life course research which was held in June, 1989 and highlighted the data analysis techniques of event history analysis and causal modeling; 2) A July 1989 workshop on methods for coding open-ended archival material for several important personality constructs; 3) A workshop on the secondary analysis of major longitudinal data sets, to be taught by Glen Elder, George Vaillant, and their associates in October, 1989; 4) A workshop on using case studies for the study of individual lives, held in May, 1990.

Grants

In order to promote actual use of archival data, we have a program of research grants. First, there is the Radcliffe Research Support Program, which offers small grants to post-doctoral researchers who wish to use data housed at the Center. In addition, there are three awards for dissertations in the areas of sex differences or female development, women's life choices and patterns, and personality or "the study of lives."

The Longitudinal Studies Inventory

A major task of the current NIMH grant has been to expand and update the longitudinal studies inventories originally published by the Social Science Research Council (Inventory of Longitudinal Studies of Middle and Old Age, Midgal, Abeles, & Sherrod, 1981, and Inventory of Longitudinal Research on Childhood and Adolescence, Verdonik & Sherrod, 1984). This has involved entering the existing inventories into a computer database, developing a detailed index, and coding the studies according to index categories. At the same time, we have been seeking out information on studies appropriate for listing but not included in the earlier volumes. In the coming year, this process will be completed, and a new volume including studies of childhood, adolescence, and adulthood will be printed and distributed.

Other

In addition to the resources already described, we are developing a set of machine-readable files that should be useful for teachers of statistics and methods courses. We also maintain a Measures File containing copies of primarily psychological instruments, and includes as well information about the measure's development, validity, reliability, and scoring. In addition, we maintain a Feminist Critique File and bibliography. This is a collection of critiques of social science methods and theory, which is updated annually.

Users

Data from the Murray Center archive are actively used for research in psychology, sociology, psychiatry, education, political science, history, economics, and criminology. Reanalyses of the data are carried out in order to explore questions that are very different from those addressed by the original investigator's analyses. Often these reanalyses are cross disciplinary, involving researchers from one discipline using data collected by investigators from another discipline. A recently published book by historian Elaine Tyler May illustrates this kind of cross-disciplinary reanalysis. The book, Homeward Bound: American Families in the Cold War Era, (May, 1988) draws very centrally on Lowell Kelly's 50-year longitudinal study of personality development within marriage, which is archived at the Murray Research Center. In this work, Professor May traces the connections between politics and family life in the 1950s and argues that the cold war affected virtually all aspects of life, from consumerism to sexuality. As Professor May notes, "The participants in the Kelly Longitudinal Study were among the cohort of Americans who began their families during the early 1940s, establishing the patterns and setting the trends that were to take hold of the nation for the next two decades. They entered

marriage when World War II thrust the nation into another major crisis, wreaking further havoc on families. They raised children as the cold war took shape, with its cloud in international tension and impending doom" (p. 12). This study of family life in the context of an important era in America history parallels the ground-breaking work on the psychological impact of the Great Depression carried out by Glen Elder with the Berkeley and Oakland longitudinal studies (Elder, 1979). Clearly, this kind of work requires archival data, including the recoding of original subject records.

Other research conducted at the Murray Center involves more quantitative reanalysis of machine-readable data. For example, political scientist Eileen McDonagh is using congressional voting records to investigate the political mechanisms through which policy innovation can occur at the national level (McDonagh, in press). The project introduces a new measure of constituency issue position — district level referenda votes — to study policy congruence between grass-root electorates and House representatives. The investigation of policy congruence processes in the context of partisan electoral patterns and demographic characteristics provides for multivariate analysis of the direct, indirect, and interactive effect of these major sources of policy formation.

Another example is provided by Janet Giele's reanalysis of several surveys of college educated women (Giele & Gilfus, in press). Among other things, Giele found that college-educated black women led college-educated white women in the shift that began to occur in the late 1950s and early 1960s toward more complex and heterogeneous life patterns.

Perhaps most ambitious are studies that involve longitudinal follow-up of an existing sample. The Murray Research Center is the only archive that makes this kind of study possible. A recent example is provided by David McClelland and his colleagues' follow-up of the Sears, Maccoby, and Levin "Patterns of Childbearing" sample (Sears, Maccoby, & Levin, 1957). First studied at age 5, and recontacted on multiple occasions in the interim, the subjects were age 41 at the time of the McClelland follow-up. Among the issues addressed by the follow-up are the family origins of empathic concern (Koestner, et al., in press), changes in motivational patterns in adulthood, the relation between agency motivation and health (McClelland, 1989), and the relation of early patterns of parental discipline and warmth to successful adaptation in adulthood (Franz, et al., 1989).

Conclusion

In conclusion, I'd like to agree that the Murray Research Center is unique, but I hope soon will not be very much less visible! I'd like to thank Sue Dodd for including us in this session toward that end.

References

Elder, G. (1979). Historical change in life patterns and personality. In P. Baltes & O. Brim (eds.)., Life-span development and behavior (Vol. 2, pp. 117-159). New York: Academic Press.

Giele, J., & Gilfus, M. (in press). Race and college differences in life patterns of educated women. In J. Antler & S. Bilken (eds.), Women and educational change. Albany: State University of New York Press.

Koestner, R., Franz, C., & Weinberger, J. (in press). Family origins of empathic concern: 26-year longitudinal study. Journal of Personality and Social Psychology.

May, E. (1988). Homeward bound: American families in the cold war era. New York: Basic Books.

McClelland D. (1989). Motivation factors in health and disease. American Psychologist, 44, 675-683.

Migdal, S., Abeles, R., & Sherrod, L. (1984). An inventory of longitudinal studies of middle and old age. New York: Social Science Research Council.

Sears, R., Maccoby, E., & Levin, H. (1957). Patterns of childbearing. Stanford University Press.

Thomas, A., & Chess, S. (1977). Temperament and development. New York: Brunner/Mazel.

Verdonik, F., & Sherrod, L. (1984). An inventory of longitudinal research on childhood and adolescence. New York: Social Science Research Council.

Distribution Of Census Data On CD-ROM To Depository Libraries

by Juri Stratford [1]
Documents Librarian
Shields Library
University of California, Davis

Introduction

The Depository Library Program (DLP) was established by Congress to inform the public on the policies and programs of the Federal government. Through the DLP, the Government Printing Office (GPO) distributes publications to designated libraries. While government documents received through depository distribution remain the property of the Federal government, depository institutions are responsible for the maintenance of, and providing public access to, the documents. This usually involves committing public service and technical service staff, and the purchase of bibliographic tools in addition to committing space.

In the past, depository libraries have provided public access to the printed Census publications while access to Census machine-readable datafiles (i.e. tapes) has been available through State data centers and data archives. The Census Bureau is now beginning to distribute machine-readable datafiles on CD-ROM to depository libraries. This affords depository libraries both new opportunities and new responsibilities.

The Depository Library Program

There are almost 1,400 depository libraries. They are located in each state and Congressional district in order to make government publications widely available. These government publications are available for the free use of the general public. For the purpose of depository distribution, a government publication is defined as "informational matter which is published as an individual document at Government expense, or as required by law." [44 USC 1901]

The origins of the DLP can be traced back to the 1790s when the State Department distributed acts of Congress to State governments and newspapers. Funding was sought on an ad hoc basis until 1813 when Congress passed a resolution authorizing "every future Congress" to print additional copies of Congressional publications for this purpose. In 1814, the American Antiquarian Society was designated the first depository library. Responsibility for the program shifted between various agencies and departments, mainly the Department of State and the Department of Interior, throughout 19th Century. Congressional resolutions in 1857 and 1858 affirmed the distribution of congressional materials to institutions such as libraries and colleges, and other organizations designated by Members of Congress. In 1895, a new printing act was passed, incorporating the old legislation and placing responsibility for the DLP in the office of the Superintendent of Documents at GPO; the act also specified that certain executive materials were to be included.[2]

The present law, the Federal Depository Act of 1962, increased the number of possible depository libraries; established a system of regional depository libraries which were to maintain a permanent collection, and provide interlibrary loan and reference service; expanded the variety of government documents available for distribution; and established a reporting mechanism, the Biennial Survey, to ascertain the libraries' condition. The 1962 Act has been amended twice: in 1972 to exempt the highest appellate court of each State from the requirement of public access; and in 1978 to extend depository eligibility to law schools.[3]

Depository Distribution of Machine-Readable Datafiles

GPO has reversed its position on depository distribution of machine-readable datafiles since the early 1980s. At their Fall 1981 meeting, the Depository Library Council (DLC), an advisory body to the Public Printer and the Superintendent of Documents, passed a resolution regarding the feasibility of the GPO providing free access for depository libraries to unclassified bibliographic data bases belonging to Federal agencies. In response to the resolution, GPO general counsel Garrett Brown determined that "...the Depository Library Act of 1962 does not direct the Superintendent of Documents to make published documents available in all possible formats to the libraries. It was the intent of Congress that only printed publications be made available to depositories."[4]

Following the Census Bureau's plans to distribute data from the 1982 Census of Agriculture and the 1982 Census of Retail Trade through the depository library program on CD-ROM as Test Disc 2, the Public Printer requested approval through the Joint Committee on Printing (JCP). In a March 25, 1988 letter to the Public Printer, Congressman Frank Annunzio, Chairman of the JCP, affirmed the Committee's support of the Census project and the GPO's authority to produce and distribute

Government publications in electronic formats. [5]

In 1989, GPO asked its General Counsel Grant G. Moy, Jr. to review the 1982 opinion. He concluded that "the earlier question presented to the General Counsel concerned only the issue of access to unpublished information in a computer data base," and, this was still the case; but "the specific statement in the General Counsel's 1982 opinion, limiting depository distribution to printed products, was disapproved."[6]

Test Disc 2 was distributed to 173 depository libraries as a pilot project in September 1988. The CD was available through regular depository distribution in 1989. At the March 1989 Depository Library Council meeting, Jan Erickson of the Government Printing Office reported on the initial distribution of Test Disc 2, and indicated that it was the Census Bureau's intent to distribute future CD-ROM products through the depository system.[7] A second Census CD, The City and County Data Book, was distributed to depository libraries in Spring 1990. Software accompanying early shipments of the City and County Data Book were "infected" with the Jerusalem virus.

Test Disc 2

The Census Bureau's CD-ROM Test Disc 2 contains data from the 1982 Census of Retail Trade and the 1982 Census of Agriculture on a single compact disk. The files are in dBase III format.

Files from the Census of Agriculture contain 1982 data by county, with comparable data for selected items from the 1978 census. The technical documentation for the Census of Agriculture describes the data as a single file with a logical record size of 40,320 characters containing 3,360 data fields. However, the dBase III record structure only allows 128 fields. This large record structure is accommodated by storing the data in 28 separate dBase III compatible files. The first file, AG82_GEO.DBF, provides geographic information for each state and county. This information is a guide to the arrangement of data contained in the 27 numeric datafiles. For example, the geographic area indicated by record #170 in AG82_GEO.DBF is "California." This means that state level data for California in the other 27 files is contained in record #170. Each file is named AG82_NN.DBF where NN = 1 to 27. Except for the last file, which contains the last 32 fields, each file contains 128 fields.

Data from the 1982 Census of Retail Trade are available for each 5-digit zip code, including number of establishments, by kind of business, and basic data for retail trade total. The files for the Census of Retail Trade have a much simpler record structure than the files for the Census of Agriculture. The 1982 Census of Retail Trade data are stored in 51 separate files. Each file is named RC82_XX.DBF where XX = the state postal abbreviation. Data files for the retail trade data have a record length of 155 characters containing 19 fields. There are two files for each state, a DBF file and an NDX file.[8]

County and City Data Book

The 1988 County and City Data Book CD-ROM contains the same data as published in the printed volume. These files contain data gathered from a variety of Federal agencies and national associations. The disc includes data for states, counties, cities with a population of 25,000 or more, and places with a population of 2,500 or more. Like the Census of Agriculture files on Test Disc 2, each record represents a geographic area, and subject fields are distributed across a number of datafiles. The datafiles range in size from about 200kb to 1mb.

The Economic Censuses and the Census of Population and Housing

Almost all data from the Economic Censuses previously available on magnetic tape will be on CD. The Economic Census and Census of Agriculture will be available on 9 CDs. The first disc, Volume 1, release 1A was released in early 1990, but has not yet been distributed to depositories. The disc contains data from the geographic area series for wholesale, retail, and service industries for selected states and includes the same statistics statistics as published in the corresponding report series: Geographic Area Series for 1987 Censuses of Retail Trade (all states), Wholesale Trade (selected states), and Service Industries (selected states); Preliminary Industry Series for the 1987 Census of Manufactures (national with some state totals) and selected historical statistics. Plans for the 1990 Census of Population and Housing call for 20-30 CDs to be released from mid- 1991-1993, including redistricting data and block statistics.[9]

Data Extraction

Each of these CD-ROM products is distributed with a program to display tables, but software is not provided to copy data subsets. The datafiles range in size from 30kb to 3mb; most of these files are too large to manipulate on a microcomputer without first creating a smaller data subset that can be copied to a floppy disk or hard disk. The texts documenting Test Disc 2 and the City and County Data Book discuss using dBase III to work with the files.

The documentation for the Economic Censuses describes the EXTRACT program. The EXTRACT program is a public domain program that was developed to create subsets from the large CD-ROM databases and save them as files on a floppy or hard disk. Extracted files can be created in dBase format, ASCII fixed field format, or ASCII comma-delimited format. Version 1 of EXTRACT is slow and occasionally crashes. A new version of EXTRACT should be released shortly. EXTRACT is

available from the Center for Electronic Analysis, University of Tennessee. However, it has not been distributed with the CD-ROMs.

Conclusion

In their paper, "Government Information in Machine-Readable Data Files: Implications for Libraries and Librarians," Ray Jones and Thomas Kinney examine the requirements for the utilization of machine-readable datafiles in retrieval and reference services. Two of their remarks can be paraphrased to apply to the utilization of the Census CD-ROMs in depository libraries. First, when librarians have the responsibility of retrieving numeric information from CD-ROMs either they must know how to program or work with a colleague who programs; and second, librarians will require the critical judgment to determine when data retrieval from CD-ROMs is needed to answer the patron's need most completely.[10] These skills are not widely held by depository librarians at the present time. This is evidenced by the fact that few depository libraries have successfully integrated these materials into their reference service.[11]

The Census Bureau is currently reviewing the impact of the CD-ROM distribution. The report, The Role of Intermediaries in the Interpretation and Dissemination of Census Data Now and in the Future, by Census statistician Sandra Rowland examines these issues. The study credits the experience of librarians assisting in the understanding and use of census data; however, it concludes that "neither the GPO nor the libraries play a big part in the interpretation of data for users" and argues that role of depository libraries is "unlikely to change in the future unless librarians take a more aggressive role as information technicians." It also states that, while the Regional Depository libraries will acquire and hold all census products including data on high density optical storage devices, most depository libraries will acquire and hold fewer census products in the future than they do now.[12]

At present, the Census Bureau appears to have a strong committment to the depository distribution of their CD-ROM products, and these materials are available to all depository libraries. The depository library community needs to work closely with data archivists to insure that effective use is made of the Census CD-ROM products. Data archivists might try to meet informally with depository librarians within their own institutions to discuss how access to Census data on tape, CD-ROM, and paper copy could best be coordinated. Data archivists might also consider coordinating presentations with state or national government document groups.

Finally, while the Census Bureau assures us that the CD-ROM production of the 1987 Economic Census and the 1990 Census of Population and Housing will not be produced at the expense of the publication of the tape or paper products, Sandra Rowland's report suggests that we can expect to see fewer paper products and more electronic products for the 2000 Census: "With respect to the year 2000 census, it is very likely that there will be a movement out of printed media and into electronic media for dissemination to the libraries."[13] While there are certainly instances where researchers' needs would best be served by Census data on CD-ROM, documents librarians and data archivists alike need to monitor the situation to insure that the CD-ROMs are not produced at the expense of other necessary Census products.

[1] Presented at the IASSIST 90 Conference held in Poughkeepsie, N.Y. May 30 - June 2, 1990.

[2] Hernon, Peter, Charles R. McClure and Gary P. Purcell, GPS's Depository Library Program: A Descriptive Analysis (Norwood, New Jersey: Ablex Publishing Company, 1985), pp.3-7.

[3] U.S. Congress, Joint Committee on Printing, A Directory of U.S. Government Depository Libraries (Washington, D.C.: U.S. Government Printing Office, 1988), pp.1-3.

[4] U.S. Congress, Joint Committee on Printing, Provision of Federal Government Publications in Electronic Format to Depository Libraries: Report of the Ad Hoc Committee on Depository Library Access to Federal Automated Data Bases... (Washington, D.C.: U.S. Government Printing Office, 1984), pp.112-13.

[5] U.S. Congress, Office of Technology Assessment, Informing the Nation: Federal Information Dissemination in an Electronic Age (Washington, D.C.: U.S. Government Printing Office, 1988), p.143.

[6] Ibid.

[7] "Summary: Spring Meeting, Depository Library Council, Pittsburg, Pennsylvania, March 8-10, 1989," Administrative Notes 10 (August 1989): 3.

[8] Stratford, Juri, review of "CD-ROM Test Disc 2 (machine-readable data file) and CD-ROM Test Disc 2: Technical Documentation," Government Publications Review, 16 (1989): 397-398; see also Peter Hernon and Candy Schwartz, "Readers Exchange: Census Product Review," Administrative Notes 11 (March 1990): 13-21.

[9] Administrative Notes 10 (August 1989): 3.

[10] Jones, Ray and Thomas Kinney, "Government Information in Machine-Readable Data Files: Implications for Libraries and Librarians," Government Publications Review, 15 (1988): 25-32.

[11] Diane Smith examines the viability of federal depositories to deal with electronic information in her forthcoming paper, "Depository Libraries in the 1990s: Whither or Wither Depositories?," Government Publications Review, 17 (1990).

[12] Rowland, Sandra, The Role of Intermediaries in the Interpretation and Dissemination of Census Data: Now and in the Future, (Washington, D.C.: U.S. Bureau of the Census, 1989), photocopy, pp.32-33.

[13] Op cited, p.14.

DATA NEWS

Here are two items that readers of the Quaterly may find of interest. (Submited by: Jim Jacobs jajcobs@ucsd.edu)[1]

1) In the newest edition of the ASIS "Annual Review of Information Science and Technology" (vol. 25, 1990, Martha E. Williams, ed., published for ASIS by Elsevier Science Publishers pp. 3 - 54), Karen J. Sy of the University of Washington and Alice Robbin of the University of Wisconsin have written an article titled: "Federal Statistcal Policies and Programs: How Good Are the Numbers?"

In their concluding remarks they say, "...the 1980s saw a serious decline in the quality of our federal statistical system." And, "Throughout the government, but especially in the key coordinating branch of the Office of Management and Budget, we found policy makers who viewed statistics as a burden and not as indicators of who we are as a nation and where we should be going. In sum, the federal statistical system is in serious trouble."

2) The Committee on National Statistics and the Social Science Research Council with support from several agencies, have convened a Panel on Confidentiality and Data Access. The scope of the study includes publicly supported statistical data collection activities. The Panel is in the middle of a two year study and is soliciting short statements from interested parties on the following topics:

Access Problems. (examples of instances where confidentiality law or policies have made it impossible to obtain data.)

Suggestions for Improving Access. (including suggestions for improving access with appropriate safeguards to maintain confidentiality.)

Persons or Businesses Harmed by Disclosure.

You can submit statements to:
George T. Duncan, chair
Panel on Confidentiality and Data Access
Committee on National Statistics
National Academy of Sciences
2101 Constitution Ave, NW
Washington DC 20418

If you have questions, or if you want a more detailed announcement of the charge to the Panel, you may call Virginia de Wolf, Study Director, (202) 334-2550.

[1] Jacobs, Jim. 1991 "Quality of Government Statistics" [computer file]. Edmontonl, Alberta: OR-L. Electronic listserv. (LISTSERV@UALTAVM).

Accessing City-County Data Book via DBASE III: Census CD-ROMs from the Ground Up

Fred Gey
Data Archivist

UC Data Archive & Technical Assistance (UC DATA)
(formerly State Data Program)
University of California
Berkeley, CA 94720

Presented to the *1990 Conference of the International Association of Social Science Information Service and Technology (IASSIST90),* Poughkeepsie, New York, May 30 to June 2, 1990.

The Bureau of the Census has published the 1988 City-County Data Book data files on a CD-ROM disk. Physically the data are organized as DBASE-III database files and the Bureau has supplied a computer program to display profiles of particular areas. However, accessing the data for analysis purposes (such as finding the ethnicity of American counties) can only be done by directly using DBASE-III, and doing so is a multi-step process. This paper describes how to use DBASE-III directly on CCDB88 to select subsets of data for statistical analysis, and compares and constrasts the CCDB88 structure with the 1982 Census of Agriculture files on Test Disk 2. Some suggestions are made for how the Bureau might organize CD-ROM data and access software to better facilitate individual access and prepare for the 1990 Census.

Accessing City-County Data Book with DBASE III

1. Introduction

In the 1990s the Bureau of the Census will utilize the CD-ROM as a major publishing medium for current and future census data. This process has been underway for several years and the Bureau has been through several iterations of test disks before settling on a de-facto standard for distribution file formats of the DBASE-III database system by Ashton-Tate Corporation. While some measure of standardization is imposed by this decision by the Bureau, other standards and capabilities need to be added for the convenience of public data users of census data. It is the purpose of this paper to discuss the kinds of tasks which might be commonplace by data users and the effort required to accomplish them, as well as to draw some conclusions as to the structure the Bureau might impose to ease the burden of accessing this data.

2. What is CD-ROM?

CD-ROM is a new data storage medium based on audio compact disk technology. Each cd-rom disk holds about 650 megabytes of data (about the equivalent of 4 computer magnetic tapes as we know them on the IBM mainframe). Since the production process is identical to that of compact disks, the cost of production is about the same (less than $5 per disk).

CD-ROM can be conceived of as an extremely large and slow disk that you can't write on. It operates like a cross between a hard disk (because it's so big) and a floppy (because the disks are removable). On the IBM-AT machine in our archive, we have attached a CD-ROM drive from Denon Corp, and designated it to be drive F:

Thus if we have the City-County Data Book disk inserted in the drive we can do a directory command to see what files are on the drive:

Directory listing of database files on CCDB CD-ROM

```
D:\FRED> dir f:*.dbf/w
 Volume in drive F is CCDB_1988
 Directory of  F:\
C1F01    DBF   C1F01DCT DBF   C1F02    DBF   C1F02DCT DBF   C1F03    DBF
C1F03DCT DBF   C1F04    DBF   C1F04DCT DBF   C1F05    DBF   C1F05DCT DBF
C1F06    DBF   C1F06DCT DBF   C1F07    DBF   C1F07DCT DBF   C1F08    DBF
C1F08DCT DBF   COF01    DBF   COF01DCT DBF   COF02    DBF   COF02DCT DBF
COF03    DBF   COF03DCT DBF   COF04    DBF   COF04DCT DBF   COF05    DBF
COF05DCT DBF   COF06    DBF   COF06DCT DBF   COF07    DBF   COF07DCT DBF
COF08    DBF   COF08DCT DBF   COF09    DBF   COF09DCT DBF   COF10    DBF
COF10DCT DBF   COF11    DBF   COF11DCT DBF   COF12    DBF   COF12DCT DBF
COF13    DBF   COF13DCT DBF   COF14    DBF   COF14DCT DBF   COF15    DBF
COF15DCT DBF   COF16    DBF   COF16DCT DBF   COF17    DBF   COF17DCT DBF
COF18    DBF   COF18DCT DBF   DCT_COU  DBF   DCT_CTY  DBF   DCT_PLC  DBF
DCT_STA  DBF   PLF01    DBF   PLF01DCT DBF   STF01    DBF   STF01DCT DBF
STF02    DBF   STF02DCT DBF   STF03    DBF   STF03DCT DBF   STF04    DBF
STF04DCT DBF   STF05    DBF   STF05DCT DBF   STF06    DBF   STF06DCT DBF
STF07    DBF   STF07DCT DBF   STF08    DBF   STF08DCT DBF   STF09    DBF
STF09DCT DBF   STF10    DBF   STF10DCT DBF   STF11    DBF   STF11DCT DBF
STF12    DBF   STF12DCT DBF   STF13    DBF   STF13DCT DBF
       84 File(s)          0 bytes free
D:FRED>
```

3. What data is available?

We currently have data from the Census Bureau, and will soon obtain a great deal more. The following data files are on hand:

Disk	File	Geography
TEST1	1980 Census, STF3	zipcode
AHS85	1985 American Housing Survey, National Core	nat/state
TEST2	1982 Census of Agriculture 1982 Census of Retail Trade	state/county zipcode
CCDB	1988 City-County Data Book	state/county/ city/place

TABLE 1 : CD-ROM Databases at UC DATA

4. Accessing CD-ROM data: City-County Data Book

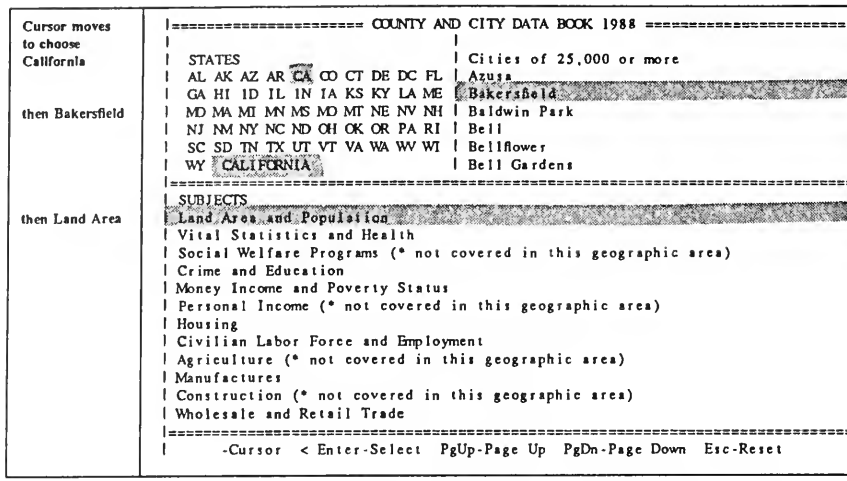
The Census Bureau has supplied some access software (in the form of canned profiles which can be applied to generate reports for particular areas), all of which is located in the CDROM subdirectory on drive C in the archive. One of these is the CCDB profile program, which can be accessed as follows:

Census Bureau CD-ROM access software

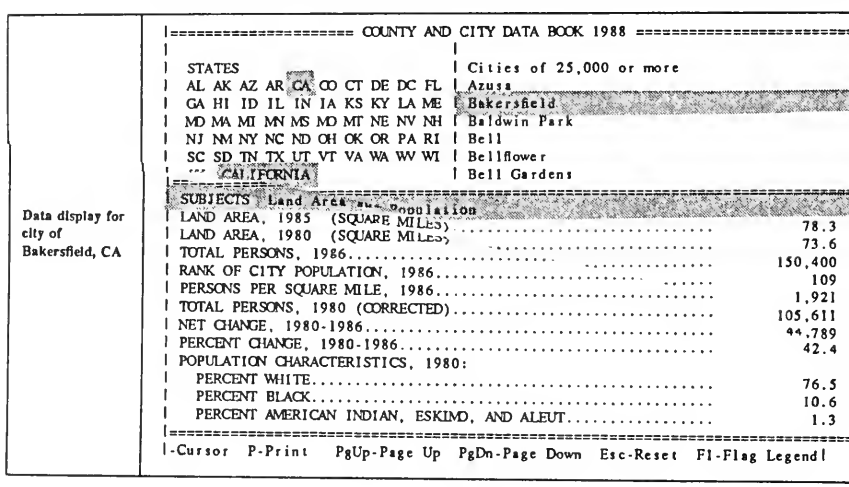
```
C:\ cd cdrom
C:\CDROM> dir

 Volume in drive C ... SDP1-130686
 Directory of  C:\CDROM

AG82          <DIR>        7-12-88   9:34a
RETAIL        <DIR>        7-12-88   9:35a
ZIPS          <DIR>        8-02-89   8:56a
README   1        666     8-08-88  10:41a
README   2        747     8-08-88  10:46a
README   BAT       58     6-23-88  10:23a
TAB36    DBF    35815     6-29-88   9:54a
AGR      EXE   179692     7-12-88   1:20p
CCDB     EXE   202788     8-24-89  11:26a
CDREADER EXE    39067     8-08-88  10:30a
RETAIL   EXE    83026     7-09-88   1:20p
DATA     SCR     4008     4-28-88   9:17a
MENU     SCR     4103     6-08-88   4:19p
CALIF    DBF      276     2-04-90   2:39p
       16 File(s)    4771840 bytes free
```

Start CCDB profile

```
C:\CDROM> ccdb
```



Choosing the city of Bakersfield and the Land Area and Population subject (as shown by shading above) will give the following data about Bakersfield:



City county data book has data for four levels of geography: states, counties within states, cities (of 25,000 population or greater), and census designated places (including unincorporated towns of 2,500 population or greater for which census data has been tabulated). The example above retrieved the first screen of items available for the city level of geography.

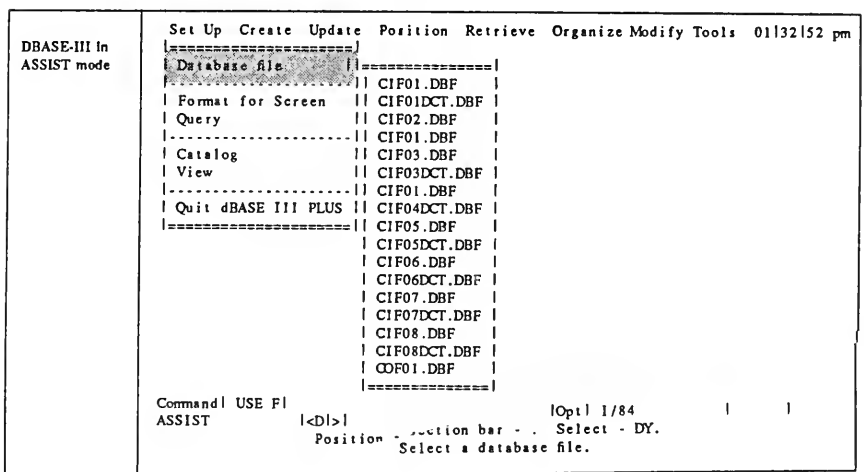
4.1. CCDB individual data

As part of the UC DATA operations and services on CCDB, the following tasks might be desired:

- Obtain a data file of all cities and towns in California containing population and per-capita income.

- Determine which counties in the United States have Hispanic population greater than 15 percent of total population, and rank them by percent Hispanic.

In order to achieve these tasks we must use DBASE-III directly. Assuming we have started DBASE correctly, and then set the drive to F, and the following screen should appear:



4.2. Choosing California Towns and Cities

In order to obtain data for towns and cities in California, we need to know (from the Census Bureau Documentation [CENSUS 89]) that the data file we are looking for is the *Place* file which contains incorporated and unincorporated (census designated) places with 2,500 or greater population in 1980. We must choose this file as our DBASE data file. This is easiest done by pressing the ESC key until we get the DBASE command prompt (the dot in column 1) and typing the "use <data base file>" command, and then using the DBASE *copy* command to actually create the new file.

Start place level database and show structure

```
use f:plf01.dbf
display structure
Structure for database: f:plf01.dbf
Number of data records:     9593
Date of last update    : 06/06/89
Field  Field Name  Type        Width    Dec
    1  STCO        Character       5
    2  MCDCODE     Character       3
    3  PLACECOD    Character       4
    4  LEVEL       Character       1
    5  AREANAME    Character      36
    6  MNY93080    Numeric         9
    7  MNY93086    Numeric         9
    8  MNY92079    Numeric         7
    9  MNY92085    Numeric         7
   10  MNY93186    Numeric         6       1
   11  MNY92185    Numeric         6       1
** Total **                      94
```

Copy California subset

```
copy to d:plf01ca.dbf for stco='06'
     401 records copied
Command Line     :<D:>:PLF01                    :Rec: EOF/9593      :       :

                        Enter a dBASE III PLUS command.
```

and then we can browse the new file to see its contents.

To browse California extract file

```
use d:plf01ca
browse
||==================|====================|==============|====================|
|| CURSOR   <-- --> |          UP   DOWN |    DELETE    | Insert Mode|  Ins  |
||  Char            | Record|  ^X   ^Y   | Char|  Del   | Exit|        End  |
||  Field| Home End | Page|  PgUp PgDn   | Field|  ^Y   | Abort|       Esc  |
||  Pan|    ^<- ^-> | Help|  F1          | Record| ^U   | Set Options| ^Home  |
||==================|====================|==============|====================|
|STPL-- LEVEL MSA- PMSA CENTRAL STATE PLACE AREANAME------------------- STATE
|060000 1                0       06    0000  CALIFORNIA                  CA
|060010 2          7362 5775 0           06    0010  Alameda               CA
|060025 2          4472 4480 0           06    0025  Alhambra             CA
|060085 2          4472 0360 1           06    0070  Anaheim              CA
|060105 2          4472 4480 0           06    0085  Antioch              CA
|060175 2          4472 4480 0           06    0105  Arcadia              CA
|060180 3          0680      1           06    0175  Azusa                CA
|060185 2          4472 4480 0           06    0185  Baldwin              CA
|060210 2          4472 4480 0           06    0210  Bell                 CA
|060215 2          4472 4480 0           06    0215  Bellflower           CA
|
|
|
|BROWSE          |<F|>|CIF01                    |Rec| 40/1008      |      |
|
|                              View and edit fields.
```

4.3. Selecting Hispanic Counties

While obtaining California places only presented a mild challenge, the process of discovering US counties with substantial Hispanic population concentration requires significantly more detective work. The Census Bureau does not make it easy because they don't include a comprehensive codebook to document the CCDB CD-ROM file, and so we must search for the data element (field in DBASE terminology) which has Hispanic population. We can begin by examining what documentation the Bureau does provide, as shown in the following table:

Table B. Counties — **Population Characteristics and Households**

County	Population characteristics—Con.												Households					
	1984—Con.									1980			1985			1980		
	Percent—									Percent—							Percent—	
	Under 5 years	5 to 14 years	15 to 24 years	25 to 34 years	35 to 44 years	45 to 54 years	55 to 64 years	65 to 74 years	75 years and over	American Indian, Eskimo, and Aleut	Asian and Pacific Islander	Hispanic[1]	Number	Percent change, 1980-1985	Persons per household	Number	Female family householder[2]	One-person[3]
	14	15	16	17	18	19	20	21	22	23	24	25	28	27	28	29	30	31

[1]Hispanic persons may be of any race. [2]No spouse present. [3]Householder living alone.

TABLE 2 : CCDB County Data Elements

the numbers above the column definitions refer to a field called *ITEM<number>* in DBASE, where <number> is replaced by the actual column number. Looking at the table, we find that percent Hispanic population is ITEM25. Unfortunately, however, we don't know which of the 18 DBF files that ITEM25 is to be found. We can guess that it's probably in COF02.DBF, and display structure for this data subset, as shown below:

structure of second county dbf file

note number of county records

```
. use f:cof02.dbf
. display structure
Structure for database: f:cof02.dbf
Number of data records:     3191
Date of last update    : 06/16/89
Field  Field Name  Type       Width     Dec
    1  STCO        Character      5
    2  LEVEL       Character      1
    3  MSA         Character      4
    4  PMSA        Character      4
    5  AREANAME    Character     36
    6  FLAG14      Numeric        1
    7  ITEM14      Numeric        6       1
    8  FLAG15      Numeric        1
    9  ITEM15      Numeric        6       1
   10  FLAG16      Numeric        1
   11  ITEM16      Numeric        6       1
   12  FLAG17      Numeric        1
   13  ITEM17      Numeric        6       1
   14  FLAG18      Numeric        1
   15  ITEM18      Numeric        6       1
   16  FLAG19      Numeric        1
Press any key to continue...
Command Line    :<D:>:COF02                    :Rec: 1/3191      :       :

                    Enter a dBASE III PLUS command.
```

But as we can see, this isn't the case.

Fortunately, the Census Bureau has not left us completely in the lurch. Several *dictionary files* have been constructed which describe the contents of the data items and where they are located in the many DBF files.

Directory listing of dictionary files

```
. dir dct*
Database Files    # Records    Last Update     Size
DCT_COU.DBF            391     06/07/89       29878
DCT_CTY.DBF            270     06/07/89       20682
DCT_PLC.DBF              6     06/06/89         494
DCT_STA.DBF            292     06/07/89       22354

  73408 bytes in     4 files.
      0 bytes remaining on drive.
```

Choose county dictionary

```
. use dct_cou.dbf
. display structure
Structure for database: F:dct_cou.dbf
Number of data records:     391
Date of last update    : 06/07/89
Field  Field Name  Type       Width    Dec
    1  ITEM        Character      8
    2  DESC        Character     57
    3  FILE        Character      8
    4  SUB_NO      Numeric        2
** Total **                     76
```

Browse the dictionary

```
. browse
Command Line     :<F:>:DCT_COU                 :Rec: 1/391          :       :

                          Enter a dBASE III PLUS command.

|=================|====================|==============|====================|
| CURSOR  <-- -->  |         UP   DOWN   |    DELETE    | Insert Mode:  Ins  |
|  Char:           | Record:             | Char:   Del  | Exit:         ^End |
|  Field: Home End | Page:  PgUp  PgDn   | Field:  ^Y   | Abort:        Esc  |
|  Pan:   ^<-  ^-> | Help:   F1          | Record: ^U   | Set Options: ^Home |
|=================|====================|==============|====================|
|ITEM---- DESC------------------------------------------------------ FILE----
|ITEM22     PERCENT 75 YEARS AND OVER                                  COF02
|FLAG22                                                                COF02
|           TOTAL POPULATION - USED FOR COMPUTING
|ITEM22A      PERCENTS FOR RACE AND AGE                                COF02
|         POPULATION CHARACTERISTICS, 1980:
|ITEM23     PERCENT AMERICAN INDIAN, ESKIMO, AND ALEUT                 COF03
|ITEM24     PERCENT ASIAN AND PACIFIC ISLANDER                         COF03
|ITEM25     PERCENT HISPANIC                                           COF03
|           TOTAL POPULATION (STF-1) - USED FOR COMPUTING
|ITEM25A      PERCENTS FOR RACE AND HISPANIC                           COF03
|         HOUSEHOLDS:
|
|
|
|BROWSE          :<F:>:DCT_COU                 :Rec: 36/391         :       :
|
|                              View and edit fields.
```

Using this dictionary file tells us what data base file to use (COF03.DBF) and which item (ITEM25) to use to search on and create our restricted file.

We can now begin the process of determining which counties had high concentrations of Hispanics according the the 1980 Census.

Open the dbf containing percent hispanic

```
use f:cof03.dbf
list stco,areaname,item25
Record#  stco  areaname                    item25
      1 00000 United States                 6.45
      2 01000 ALABAMA                       0.86
      3 01001 Autauga, AL                   1.13
      4 01003 Baldwin, AL                   1.01
      5 01005 Barbour, AL                   1.08
      6 01007 Bibb, AL                      1.12
      7 01009 Blount, AL                    0.56
      8 01011 Bullock, AL                   1.51
      9 01013 Butler, AL                    1.33
     10 01015 Calhoun, AL                   1.11
     11 01017 Chambers, AL                  0.91
     12 01019 Cherokee, AL                  0.53
     13 01021 Chilton, AL                   0.55
     14 01023 Choctaw, AL                   0.97
     15 01025 Clarke, AL
Command Line    :<D:>:COF03          :Rec: EOF/3191      :       :

                    Enter a dBASE III PLUS command.
```

Copy to a new file if more than 15 percent hispanic

```
copy to hispanic.dbf fields stco,level,msa,pmsa,areaname,item23,item24,
item25, item25a for item25>15.0
    197 records copied
```

Once we have obtained this data file we must sort in descending order using the DBASE sort command, and then we can list the contents to find the highest Hispanic concentrated counties in the United States.

Use DBASE sort command to sort to a new file in descending order

```
sort to hispsrt.dbf on item25/d
  100% Sorted          197 Records sorted
use hispsrt.dbf
list stco,areaname,item25a,item25,item24,item23

Record#  stco  areaname                  item25a  item25  item24  item23
      1 48427 Starr, TX                   27266   96.93    0.05    0.10
      2 48479 Webb, TX                    99258   91.52    0.09    0.11
      3 48247 Jim Hogg, TX                 5168   90.54    0.02    0.00
      4 48323 Maverick, TX                31398   90.34    0.11    2.35
      5 48507 Zavala, TX                  11666   89.03    0.03    0.08
      6 35033 Mora, NM                     4205   86.56    0.05    0.17
      7 48047 Brooks, TX                   8428   85.99    0.04    0.06
      8 48131 Duval, TX                   12517   85.76    0.0
Command Line    :<D:>:HISPSRT       :Rec: 1/197       :       :

                    Enter a dBASE III PLUS command.
```

5. 1982 Census of Agriculture

A task which one might wish to undertake is to draw information from the 1982 Census of Agriculture for counties with high Hispanic concentrations and combine it with the County data book information just obtained. Some of this information can be found in CCDB itself, but more detailed information would lead us to Census Bureau's CD-ROM TEST DISK 2 which contains the entire 1982 Census of Agriculture. If we pop this disk into our CD-ROM

drive, we can see how the files are organized.

Directory of ag82 data files on Test Disk 2

```
display files like ag*.dbf
AG82_01.DBF     AG82_02.DBF     AG82_03.DBF     AG82_04.DBF
AG82_05.DBF     AG82_06.DBF     AG82_07.DBF     AG82_08.DBF
AG82_09.DBF     AG82_10.DBF     AG82_11.DBF     AG82_12.DBF
AG82_13.DBF     AG82_14.DBF     AG82_15.DBF     AG82_16.DBF
AG82_17.DBF     AG82_18.DBF     AG82_19.DBF     AG82_20.DBF
AG82_21.DBF     AG82_22.DBF     AG82_23.DBF     AG82_24.DBF
AG82_25.DBF     AG82_26.DBF     AG82_27.DBF     AG82_DOC.DBF
AG82_GEO.DBF

128747612 bytes in    29 files.
      0 bytes remaining on drive.

Command Line    :<F:>:                    :                :      :

                        Enter a dBASE III PLUS command.
```

Open the first dbf file and examine its structure

```
use ag82_01.dbf
display structure
Structure for database: F:ag82_01.dbf
Number of data records:     3177
Date of last update   : 02/03/88
Field  Field Name  Type       Width    Dec
    1  TAB_01_001  Numeric        12
    2  TAB_01_002  Numeric        12
    3  TAB_01_003  Numeric        12
    4  TAB_01_004  Numeric        12
    5  TAB_01_005  Numeric        12
    .
    .
    .
Press any key to continue...
  122  TAB_02_010  Numeric        12
  123  TAB_02_011  Numeric        12
  124  TAB_02_012  Numeric        12
  125  TAB_02_013  Numeric        12
  126  TAB_02_014  Numeric        12
  127  TAB_02_015  Numeric        12
  128  TAB_02_016  Numeric        12
** Total **                    1537
Command Line    :<F:>:AG82_01              :Rec: 1/3177      :      :
                        Enter a dBASE III PLUS command.
```

What we find is that the Agriculture data files, unlike the CCDB data files, *don't have any geographic codes in their records!* The sole place for the geographic codes is in a separate file AG82_GEO.DBF. What this means to the unwary analyst is that the DBASE command JOIN can't be used to merge files from these two databases; a substantially more complex program will have to be constructed if we wish to put together data from these two sources, even though they are ostensibly collected for the same counties. We can look at this geographic file and find out what it contains.

Browse the geographic reference dbf file for 1982 Census of Agriculture

```
use ag82_geo.dbf
browse
||==================|======================|==============|====================|
|| CURSOR   <-- --> |          UP   DOWN   |    DELETE    | Insert Mode:  Ins  |
||  Char:           | Record:              | Char:   Del  | Exit:          ^End|
||  Field: Home End | Page:  PgUp  PgDn    | Field:  ^Y   | Abort:         Esc |
||  Pan:   ^<-  ^-> | Help:  F1            | Record: ^U   | Set Options: ^Home |
||==================|======================|==============|====================|
STATE COUNTY NAME----------------------------------
01    000    ALABAMA
01    001    AUTAUGA COUNTY
01    003    BALDWIN COUNTY
01    005    BARBOUR COUNTY
01    007    BIBB COUNTY
01    009    BLOUNT COUNTY
01    011    BULLOCK COUNTY
01    013    BUTLER COUNTY
01    015    CALHOUN COUNTY
01    017    CHAMBERS COUNTY
01    019    CHEROKEE COUNTY
```

Note number of counties

```
BROWSE          :<F:>:AG82_GEO                  :Rec: 1/3177
                        View and edit fields.
```

Notice that the geographic codes don't have the same name for the two databases. State and county codes are separate on the 1982 Agriculture geography file, while they are concatenated into a single STCO code in the CCDB database files. This further complicates the task of merging the two files. Finally, the above screen shows 3177 counties in the data file, while looking at the CCDB screen at the bottom of page 6 shows 3191 counties in the CCDB file. While this difference can be explained by the independent cities in Virginia and by the Bureau's not releasing Agricultural data for counties with fewer than ten farms, these facts further complicate the database compatibility problem.

Thus our preliminary investigation shows merging City-County Data Book with 1982 Census of Agriculture to be a complicated process beyond the scope of this paper. The key to making this process easier will be the development of standards for geographic coding.

6. Conclusions and Recommendations

While the Census Bureau has gone a long way toward making census data available on inexpensive personal computers, certain additional features will make accessibility of CD-ROM census data to the average planner or statistician using these data.

The Census Bureau, in constructing CD-ROM products, should provide

- A comprehensive codebook for data dictionaries which not only names and describes each data item, but also gives its universe, so items are not inadvertantly combined.

  a) An effort should be made to combine items having the same universe into the same DBASE file.

- Uniformity of file structure across databases:

  a) one record per geographic area

  b) The same geographic units over all databases (e.g. either one file for all counties in the U.S. or one file per state)

  c) The same geographic naming structure within each file and across databases (e.g. STCO as in City-County Data Book or STATE, COUNTY as in 1982 Census of Agriculture). This is so that the DBASE 'JOIN' command can be used to connect data from different files or the 'LOCATE FOR' commands will use the same selection sequence for files being connected.

This purpose of this paper has been to give a glimpse of the effort necessary to do more than trivial tasks in retrieving data from the City-County Data Book on CD-ROM. In doing so we have uncovered some of the issues in access to census data. The availability of 1990 Census data on CD-ROM will surely force social science information specialists to confront these issues in this new medium. The application of standards will resolve some problems, but others can be relieved with additional machine-readable documentation from the Census Bureau.

The Merit Networking Seminars
MAKING YOUR NSFNET CONNECTION COUNT

Merit/NSFNET Information Services is committed to providing current information on national networking to all users of the NSFNET backbone. Toward this end we will sponsor a two-day seminar in Ann Arbor, Michigan, May 20 and 21. "Making Your NSFNET Connection Count" will be an informative seminar focusing on issues of interest to campus computing leaders, information systems and networking administrators, educational liaisons, librarians, and educators who want to learn more about national networking.

Day 1, "Real People Doing Real Things," will feature a numberof presentations concerning network applications in education from the elementary grades through the college level. The Day 1 activities will begin with a keynote address by Paul Evans Peters, Director, Coalition for Networked Information and will close with a tour of the Merit Network Operations Center.

Day 2, "How to Get Connected and Stay Connected" will provide local, state, mid-level, and national networking perspectives from the experts. Day 2 will also be comprised of presentations on internetworking, information/user services, and network operations.

The seminar will be held at the Tenneco Automotive Training and Development Center in Ann Arbor. Microcomputers connected to regional and national networks will be available on-site so that attendees may access network resources discussed in the presentations.

The registration fee is $395. An early-bird fee of $345 will be charged for registrations received before April 15, 1991. This fee includes the two-day seminar, a reception on Sunday evening, lunch on Monday and Tuesday, all seminar material, and an optional tour of the Network Operations Center.

For further information send an electronic message to seminar@merit.edu or telephone 1-800-66-MERIT.

3480 cartridges.

For those interested in 3480 cartridges:

NTIS (PB8 233135) is selling for $12.95 National Archives Technical Information Information Paper No. 4 "3480 Class Tape Cartridge Drives and Archival Tape Storage: Technology Assessment Report."

Call 703/487-4600 or write to Document Sales NTIS, Springfield, VA 22161.

This paper covers the "mechanical & technological future of the systems" & "provides valuable information to data center managers data librarians, and archivists, in fact to all who are concerned about the long-term storage of machine-readable data."

Sex on the Racks: Issues of Data Collection and Access

by Daniel C. Tsang [1]
Machine-Readable Data Files Librarian
Main Library
University of California, Irvine
California

In the midst of the AIDS crisis, researchers seeking accurate empirical data about the prevalence of high-risk sexual behaviors, or the population of homosexuals in the United States, are increasingly frustrated at the lack of reliable and accurate data. Instead, researchers are reaching back some forty years, relying on Kinsey-era non-generalisable data to estimate the number of homosexuals expected to come down with AIDS (Fay et al, 1989, 243).

This absence of reliable new data (except for a 1970 national study) since Alfred Kinsey's landmark studies of male and female human sexuality in the 1940s and 1950s is due to many factors, including long-standing taboos over certain sexual practices as well as political opposition. Most recently, this led to Congress scuttling, last year, a planned national survey of sexual habits of Americans, after conservative Congressmen, such as California's William E. Dannemeyer (who said the survey was "more apropos for the pages of a pornographic magazine"), mounted a successful campaign to oppose federal funding of a National Opinion Research Center (NORC) national sex survey (Associated Press, 1989; Booth, 1989a, 1989b; Dannemeyer, 1989; Hayden, 1989; "Kinsey II," 1989; Peterson, 1989; Specter, 1989, 1990).

On the other hand, the U.S. Census Bureau, in its 1990 decennial census, has been able to gather information on domestic partnerships, so that for the first time in U.S. history, lesbian and gay couples who live with each other are being counted in a national census (Vobejda, 1990). That this unprecedented exercise in data collection will not be 100% successful is apparent from the concerns some gay activists have voiced about whether they trust the government to maintain confidentiality of the data; the memory of what happened in World War II, with the release of census data to the military regarding the distribution of Japanese Americans, and their subsequent internment, remains too real to many Americans.

Despite the national setback regarding a sex survey, the Centers for Disease Control is attempting to gather data at the municipal level; a number of cities have been slated for projects assessing behaviors that are considered high risk for AIDS, although not without having to overcome local opposition from residents who fear a repeat of the Tuskegee Experiment, when the Public Health Service deliberately did not treat hundreds of black men for syphilis (Boffey, 1987; Boodman, 1988a and 1988b).

Ironically while Congress has focused its attention on stopping the national sex survey, it has overlooked the invasion of law enforcement agents into an arena traditionally the domain of social scientists, and allowed the proliferation of sex surveys aimed, not at promoting public health, but at entrapping those suspected of being interested in pornography. Hundreds of Americans have been sent to prison, in part because they filled out a bogus sex survey commissioned, surreptitiously, by U.S. Customs or the U.S. Postal Inspection Service. Our sex police have, in fact, mastered desktop publishing, and sent questionnaires to thousands of Americans, asking about intimate details of their sex lives, including whether they are interested in sex with children or with animals (see Appendix A for an example). Respondents — who may well have been indulging in taboo fantasies — are subsequently sold child pornography published by these law enforcement agencies — and after their homes or businesses are raided, sent to jail for 10 years or more for receiving pornography. The questionnaires they have filled out prevent them from claiming entrapment — because their answers on the questionnaires indicate they are predisposed to the crime of receiving pornography (Bull, 1987; Johansen, 1988; Lee, 1987; Stanley, 1988, 1989; Tsang 1987).

It is thus not surprising that, while law enforcement agents can conduct these surveys without any oversight by Congress or by human subjects review boards, independent researchers — not connected to the criminal justice establishment — are finding out that certain sexual practices — such as childhood sexuality — are taboo and cannot be researched without law enforcement involvement. In fact, an increasing number of researchers have themselves been arrested, their research confiscated, their careers destroyed, all because they picked a subject too taboo to research (Sonenschein, 1987). Furthermore, as sexologist John Money has suggested, "The only way a researcher can get Government funding is to be against sex" (Money, 1990).

Sex data collection, therefore, is a highly politicized

endeavor, especially in the post-Meese Commission era. It raises important ethical issues — not only about the ethics of breaching confidentiality (as in data on sexual partners) — but also the ethics of hiding the true (law enforcement) purpose of a sex survey (Tsang, 1989).

Politics aside, all sexual behavior studies are faced with two major technical difficulties, viz., bias in the selection of subjects and the reliability of the responses (Forman and Chilvers, 1989, 1140). Kinsey and his colleagues did not conduct a national sample, but instead relied on participants on selected campuses and in specific organizations. With homosexual acts still criminalized in half the United States and discrimination against persons with AIDS rampant, respondents may be unlikely to admit to illegal activities or tell the truth (Wolpert, 1989). Further the specter of Big Brother asking these questions and the inability to convince everyone of the confidentiality of one's responses makes the reliability of the data particularly questionable. Missing data will be a common result (Reinisch, 1988). An undercount is also the likely result.

For the data librarian or the data user, this means that one should approach these data with more than the usual skepticism. It may also mean that researchers are less willing to part with these data.

The universe of sharable existing data on human sexual behavior is rather small. First, there is very little tradition among sex researchers (like other researchers) of citing the availability of their data in their research. Although the National Science Foundation has now mandated that grantees deposit or make available their data, this has not caught on in the sex research community. Furthermore, much of what passes for empirical sex research is either of the case study variety, or college sophomores' reports of their sex habits. This type of data may be of less interest to other researchers than something more systematically gathered. With the increase in good AIDS data, however, more interest in sharing data can be expected. If the data is publicly supported, one can well argue that they should not be made inaccessible. Privately supported data will be harder to get, of course; it took almost two decades before a Kinsey Institute study of sexual behavior, conducted by NORC in 1970, was finally released to other researchers, in part because of a dispute over which researcher would be listed as the primary author (Booth, 1988; Reinisch, 1988; Klassen, 1988).

In the United States, excluding funding agencies such as the federal government, there are two main sources that collect and distribute datasets that have material relating to this topic.

The Inter-university Consortium for Political and Social Research (P.O. Box 1248, Ann Arbor MI 48106; (313) 763-5010) is the major source for much social science empirical data, and thus it would surprise no one that in fact, in many of the datasets archived at the consortium, there are data on sexual attitudes and behaviors. The ICPSR collection is now accessible in a number of ways, by looking up a keyword in the annual ICPSR Guide To Resources Subject Index (on CDNet or printed from tape), or by searching RLIN, the Research Libraries Group's cataloging database. The 1989/90 subject index lists 38 studies under the keyword "Sexual," two studies under the keyword "Sexually," two more under the keyword "Homosexuality," and three under the keyword "Homosexuals." In addition, 11 studies are listed under "AIDS."

RLIN's MDF subfile (for machine-readable data files) is perhaps a better source since each ICPSR (and non-ICPSR study) in the database is fully analyzed by subject, and one can productively search under the subjects AIDS, homosexuality, homosexuals, lesbians, sexual attitudes, rape, or sex offenders. The bulk of the sex data files listed in RLIN are bibliographic files from the WestLaw database concerning civil rights laws; however, dozens of ICPSR datasets also show up, most of those concerning sexual attitudes, and not behavior. Among recent datasets distributed by ICPSR on sexual behavior are the National Lesbian Health Care Survey, 1984-1985 (ICPSR Study 8991) and Dangerous Sex Offenders (ICPSR Study 8985); NORC's annual General Social Survey, distributed by ICPSR, also now includes questions on sexual behavior, and the 1988 General Social Survey was used in the recent analysis of the 1970 Kinsey Institute data (Fay et al, 1989).

ICPSR is also the site of the Midwest AIDS Biobehavioral Research Center, an NIH-funded project to collect survey questionnaires used in AIDS research. Thus far, some 5,000 questions have been collected, from over 40 surveys, in the hope of providing a database of questions so that some standardization and comparison studies will take place (Michael Traugott, personal communication, 25 May 1990).

A second major source of data is the Data Archive on Adolescent Pregnancy and Pregnancy Prevention (from Sociometrics Corporation, 170 State St., #260, Los Altos CA 94022-2812; (415) 949-3832), now available in part on CD-ROM. Its 1990 Catalog of Products listed only six studies under the keyword "Sexual" and two other studies under keyword "Sexuality." However, a subject search of its database, under the topic "Sexuality," produced an 88-page printout, listing descriptions and variables for 28 studies. Among them: The 1983 Cuyahoga County, Ohio, Familial Communication and Adolescent Sexual Behavior Project, with 940 variables, including preteens and teens answering questions about homosexuality, masturbation, and oral sex (Study A1-

A2). Another study focuses on sexual behavior of minority teens and preteens (Project Redirection, Study 91-94).

In addition, a new National Archive on Child Abuse and Neglect at Cornell University operated by its Family Life Development Center (E200 MVR Hall, Ithaca NY 14853-4401; (607) 255-7794), but physically located within the facilities of the Cornell Institute for Social and Economic Research, is a source for a growing number of data files on child sexual abuse. The archive is funded by a grant form the National Center on Child Abuse and Neglect.

A list of existing or proposed AIDS- or HIV-related data files appears as appendix B (four pages) in the U.S. Office of Science and Technology's 1988 report, A National Effort to Model AIDS Epidemiology. The principal investigators are identified, as are the cities or institutions where the data are based. Among the datasets named are AIDS Surveillance (Centers for Disease Control), NORC's General Social Survey, and an AIDS Behavioral Research Clearing House at Temple University.

With AIDS already having claimed the lives of more people in the United States than the number of Americans killed in the Vietnam War, a sense of urgency pervades recent calls for the establishment of a national AIDS or HIV database. The previously cited report from the Office of Science and Technology Policy, which advises the U.S. President, specifically called for the creation of a directory of relevant AIDS databases, support for access to significant local databases, and enhanced public access to national databanks. Researchers are unwilling to release data they themselves are still analyzing, and public-use AIDS case data is not released by city but rather only in terms of six larger geographical regions in the U.S. (see also Layne et al, 1988, 511). It also called for the creation and adoption of standards and guidelines for data collection, documentation and release.

In an appendix on long-term prospects for data management, the report argued against the concept that sharing of information implies the centralization of data. It elaborated as follows:

> — Data from distinct sources are often not suitable for pooling because they were not collected under similar circumstances.
> — Data from different studies require different protection measures.
> — Quality control is best exercised by the people who have the original responsibility and authority over the data contents and collection.
> — Longitudinal studies require dynamic updating; remote compilations are unlikely to remain consistent.
> — Flexibility to incorporate new data elements, as they appear to be useful, is difficult to achieve in central repositories dealing with many sources.
> — Technology and tradeoffs of systems versus personnel costs are moving computing toward distributed paradigms.

The report envisions a decentralized network-based hypertext-formatted information sharing system, best illustrated by this example of an end user sitting in front of a computer:

> At the information mode, one may locate the title of a publication. Clicking on the title can produce the abstract, kept at that node. A click on the abstract can cause the text of the paper to be fetched, and the section titles of the paper will be displayed. Clicking a section name will obtain the corresponding section of the paper. Clicking a graph can produce the underlying values. A numerical result can be clicked on to show the algorithm or program used to obtain the result, from the workstation where the computation was performed. Clicking on another marker corresponding to the data will obtain the data, subject to privacy constraints, for display on the screen. Clicking a reference cited can continue this browsing process. Text so structured may also be annotated for further private of public use (p. 63).

Although the report did not call for the establishment of a national AIDS data center, contrary to a story in The Chronicle of Higher Education (Turner, 1989), researchers involved in the report did separately call for such a center (Turner, 1989; Layne et al, 1988). As envisioned, such a center would house a national HIV database, "in its most complete form, a storehouse of raw data on HIV infection and the AIDS epidemic" (Layne et al, 1988, 512; see also Hirons et al, 1989). Proponents also propose that the database would furnish a standard agreement governing procedures on sharing raw data. They note that the creation of a national HIV database would require an "extraordinary level of commitment on the part of the research community. Individual researchers and institutions will have to share and protect large quantities of confidential data on the intimate behaviour of individuals. They will also have to share data that could otherwise be hoarded to build their own careers. But such a database is needed — and it is needed soon" (Layne et al, 1988, 512).

Supporters of the idea of a national center argue that the "current lack of a national AIDS data base center to collect, analyze and distribute the available data is a severe block to our understanding [of AIDS transmission]." As researchers who use mathematical models to understand the AIDS epidemic, they believe establishing a center "will encourage closer collaborations between modellers and data collectors" (Hyman and Stanley, 1987, VIII.3).

Given the taboo concerning sex research, the United States, not surprisingly, appears to lag behind other countries in the area of data collection and access. The World Health Organization has been at the forefront of collecting global data on sexual practices, as part of a multinational study of AIDS (Booth, 1989). In addition to Geneva, where WHO is based, some of the data will be archived at Essex, England, at the ESRC [Economic and Social Research Council] Data Archive, where the coordinator of WHO's Homosexual Response Studies is situated (APM Coxon, personal communication, 27 May, 1990). Essex has also been the site of the computerized AIDS Register, which with funding from the Medical Research Council, lists ongoing research on AIDS.

In summary, improved data collection and access to sex data will only occur if sufficient funding, political support, public trust and researcher commitment all materialize. Otherwise, sex on the racks will more likely be found in some sex club, and not in a data archive.

References

Associated Press, 1989. "Sullivan Orders Changes in Sex-Survey Questionnaire," The Orange County Register (8 April), A10.

Boffey, Philip M., 1987. "U.S. to Test for AIDS in 30 Cities; Household Sampling Put Off," The New York Times (3 December), 10.

Boodman, Sandra G., 1988a. "AIDS Study to Involve 800 D.C. Households: City Officials Say They Were Not Consulted on Federal Project," The Washington Post (28 July), A1, A18.

Boodman, Sandra G., 1988b. "Federal AIDS Study in D.C. Postponed: City Officials Say Household Survey Would Have Been Unfair," The Washington Post (29 July), A1, A12.

Booth, William, 1988. "The Long, Lost Survey on Sex," Science 239:4844 (4 March), 1084-1085.

Booth, William, 1989a. "Asking America About its Sex Life," Science 243:4889 (20 January), 304.

Booth, William, 1989b. "U.S. Probe Meets Resistance," Science 244:4903 (28 April), 419.

Booth, William, 1989c. "WHO Seeks Global Data on Sexual Practices," Science 244:4903 (28 April), 418-419.

Bull, Chris, 1987. "Feds Nab 150 in Porn Sting," Gay Community News (8-14 November), 1, 12.

Dannemeyer, William E., 1989. "Proposed 'Sex Survey'," Science 244:4912 (30 June), 1530. (Letter.)

Fay, Robert E. et al, 1989. "Prevalence and Patterns of Same- Gender Sexual Contact Among Men," Science 243:4889 (20 January), 338-348.

Forman, David, and Clair Chilvers, 1989. "Sexual Behaviour of Young and Middle Aged Men in England and Wales," British Medical Journal 298 (29 April), 1137-1142.

Hayden, Tom, 1989. "'Magic Bullets' and Deadly Taboos: What We Don't Want to Know of Sexual Behavior May Kill Us," Los Angeles Times (7 June), II, 13.

Hirons, G. et al, 1989. "An Interactive Relational Database for HIV and the Immune System," in R. A. Morrisset, ed., Ve Conference Internationale sur Le SIDA: Le Defi Scientifique et Social: V International Conference on AIDS: The Scientific and Social Challenge: Montreal, Quebec, Canada, June 4-9, 1989 (Ottawa, Ont.: International Development Research Centre), 652. (Abstract.)

Hyman, James M. and E. Ann Stanley, 1987. "Using Mathematical Models to Understand the AIDS Epidemic." Paper presented at the Los Alamos Center For Nonlinear Studies Conference on Nonlinearity in Biology and Medicine, May 18-22. Revised version in Mathematical Biosciences 90:1-2 (July/August 1988), 415-473.

Johansen, Bruce E., 1988. "The Meese Police on Porn Patrol," The Progressive (June), 20-21.

"Kinsey II," The Orange County Register (21 March 1989), B6. (Editorial.)

Klassen, Albert D., 1988. "'Lost' Sex Survey," Science 240:4851 (22 April), 375-376. (Letter.)

Layne, Scott P. et al, 1988. "The Need for National HIV Databases," Nature 333 (9 June), 511-512.

Lee, Kevin, 1987. "Sex Research or Law Enforcement?" NAMBLA Bulletin, " 8 (April/May), 3-4.

Money, John, 1990. "Sex: The Good, the Bad and the Kinky," Playboy (July), 46-49.

Peterson, Larry, 1989. "Dannemeyer Fights Proposed Sex Study," The Orange County Register (18 March), B1, B7.

Reinisch, June Machover, 1988. "Kinsey Sex Surveys," Science 240:4854 (13 May), 867. (Letter.)

Sonenschein, David, 1987. "On Having One's Research Seized," The Journal of Sex Research 23:3 (August),408-414.

Specter, Michael, 1989. "Funds for Sex Survey Blocked by House Panel: AIDS Researchers Say Data is Essential," The Washington Post (26 July), A3.

Specter, Michael, 1990. "What's America Doing in

Bed? We Need a National Sex Survey to Fight AIDS Effectively, So Why is Congress Ducking it?" The Washington Post (25 February), B1.

Stanley, Lawrence A., 1988. "The Child-Pornography Myth," Playboy (September), 41-44.

Stanley, Lawrence A., 1989. "The Child Porn Myth," Cardozo Arts & Entertainment Law Journal 7:2, 295-358.

Tsang, Daniel C., 1987. "Moral Panic in North America: Implications for Sex Professionals." Paper presented at the Annual Conference of the Society for the Scientific Study of Sex, Western Region, 27-29 March, Beverly Hills, California.

Tsang, Daniel C., 1989. "Ethical Dilemmas in Sex Research and Therapy." Paper presented at the Annual Conference of the Society for the Scientific Study of Sex, Western Region, Marina Del Rey, California, 22-25 March.

Turner, Judith Axler, 1989. "Creation of $6-Million National Center to Collect and Analyze Data on Spread of AIDS is Urged," The Chronicle of Higher Education (25 January), A4.

U.S. Office of Science and Technology Policy, 1988. A National Effort to Model AIDS Epidemiology. Washington, D.C.: Office of Science and Technology Policy.

Vobejda, Barbara, 1990. "'Unmarried Partner' Category to Provide First Census Data on Gay," The Washington Post (11 March). A6-A7.

Wolpert, Stuart, 1989. "Shaking Down the AIDS Data: People Don't Always Tell the Truth About Their Sexual Habits," UCLA Magazine (Spring), 13-14.

Appendix A
Sample questionnaire from a U.S. Postal Inspection Service sting operation in the 1980s.

[1] Paper presented at the 16th Annual Conference of the International Association for Social Science Information Service and Technology (IASSIST), Poughkeepsie, New York, May 30-June 2, 1990.

Sample questionnaire from a U.S. Postal Inspection Service sting operation in the 1980s.

THE AMERICAN HEDONIST SOCIETY
P.O. BOX 2098
MADISON, WI 53701

The AMERICAN HEDONIST SOCIETY is a private, members only society for those who adhere to the doctrine that pleasure and happiness is the sole good in life. We believe that we have the right to read what we desire, the right to discuss similar interests with those who share our philosophy, and finally that we have the right to seek pleasure without restrictions being placed on us by an outdated puritan morality.

If you believe in our doctrine and wish to be considered for membership in the AMERICAN HEDONIST SOCIETY, then please complete the application which follows. Upon receipt of your application, it will be reviewed by our membership committee. Because we are a members only society you will be notified if your application for membership is accepted by the committee. Once enrolled as a Society member you will be entitled to receive the Society's newsletter, IT'S A SMALL WORLD, (published 4 times per year), you may freely correspond or meet with others who share your interests and have been screened by the membership committee as being true hedonists and trustworthy individuals.

The yearly membership fee for the Society is only $4.00 per year which is to cover postage and printing. HOWEVER DO NOT SEND ANY MONEY AT THIS TIME. WE MAY ONLY ACCEPT A MEMBERSHIP FEE FROM THOSE SELECTED AS MEMBERS IN THE SOCIETY. WE INTEND TO RETAIN A MEMBERS ONLY POLICY AND WILL THEREFORE NOT ACCEPT MONETARY CONTRIBUTIONS FROM NON-MEMBERS. FAILURE TO COMPLY WITH THIS PROVISION COULD RESULT IN YOUR APPLICATION BEING REJECTED.

I hope you will take the time to complete and return the following Application for Membership.

APPLICATION FOR MEMBERSHIP TO THE AMERICAN HEDONIST SOCIETY

The following information is provided in considering me for membership in the AMERICAN HEDONIST SOCIETY:

A. SEX: _____Male _____Female

B. Age: _____

C. Filing for membership as a _____Single
_____Couple

D. What is your opinion regarding each of the following sexual activities:
I = in favor of the activity
O = opposed to the activity
U = undecided

_____ Heterosexual
_____ Homosexual / Lesbian
_____ Swinging (Group / Roman Sex)
_____ Swinging (Private)
_____ Pedophilia
_____ French Culture
_____ Nudism
_____ English Culture
_____ Family Love
_____ Bondage and Discipline
_____ Sadism & Masochism _____Sadist _____Masochist
_____ Water Sports
_____ Animal Training
_____ Dominance _____Dominant _____Submissive
_____ Voyeurism
_____ Greek Culture

E. What Types of sexual material do you enjoy:
1 = I really enjoy this material
2 = I enjoy this material
3 = I somewhat enjoy this material
4 = I do not enjoy this material

_____ Nude or Pinup Type
_____ Group Sex (2 or more couples)
_____ Heterosexual (1 man / 1 woman)
_____ Lesbian Sex
_____ Transvestite Material
_____ Pre-teen Sex
_____ Water Sports
_____ Masturbation
_____ Homosexual
_____ Bondage and Discipline
_____ Sadism and Masochism
_____ Transsexuals
_____ Animals
_____ 2 Man / 1 Woman
_____ 2 Women / 1 Man

Using New Technologies to Provide Easy Access to Research Databases

by Andy Covell [1]
Manager, Research Data Center
Syracuse University

The Research Data Center, a unit within Syracuse University's central computer services organization, provides fee-based programming and data management services for researchers engaged in data-intensive research. Research Data Center analysts routinely develop strategies for managing, processing, and analyzing research databases. Mainframe based access to magnetic tape and disk is the prevailing strategy for providing access to large research databases. With recent technological developments a number of alternatives can now be realistically considered, and the Research Data Center is investigating these alternatives. This paper summarizes the information obtained over the last 6 months of that investigation.

Introduction

Most research databases are created through the collection, entry and organization of data specifically for research (e.g. survey research) or are derived from data collected outside the research enterprise for reasons other than academic research (e.g. government databases).

There are three basic types of research databases:

- *Raw files* are electronically readable files which are not formatted for any particular software package so they cannot be analyzed directly. Raw files are accessed very infrequently, and are usually stored off-line once they are read into a master file.

- *Master files* are data files which have been formatted for some software package (e.g. SAS) so they can be easily accessed for direct analysis or to obtain extracts. They are usually static, and they are typically accessed on a regular basis by some community of researchers over an extended period. Storing master files off-line is common, although on-line access is clearly preferable.

- *Analysis files* are data files created for a specific research analysis. They are formatted for some software package and contain the derived variables needed to answer a specific research question. Analysis files, the working data sets of a research project, are created, modified, and analyzed with great frequency during analysis. They are rearely accessed once the research is completed. Researchers always prefer on-line access to the analysis files which support current research, while the analysis files of previous research can be stored off-line.

Access to large research databases (raw files, master files, and even analysis files) is often limited - commonly access is through mainframe attached magnetic tape drives. Recent technological developments can significantly enhance access to large databases. High speed networks, an ever-increasing array of on-line and near-line storage alternatives, and user friendly, flexible database software are rapidly evolving technologies that can provide fast easy access to large research databases.

Networks

High speed networks offer fast access to data stored on remote computers. With a few simple commands researchers can retrieve data from a remote computer, even if it is from a different model computer across the country. Research databases no longer need to be written out to magnetic tape to be transferred from one computer to another.

Over the past five years high speed networks have become a fact of life at US universities. Most have spent considerable sums on high speed campus networks, and millions have been spent on regional networks which link campus nets together and a national backbone which links the regionals. One result is the Internet, a large conglomeration of interconnected campus, regional and national networks.

Basic, consistently implemented network services—remote login, file transfer, and electronic mail—are available on all Internet computers (except some desktop computers which may lack electronic mail). Researchers use the same procedure to access a file whether the file is on a computer located in another building on campus or on an Internet computer located in another part of the country.

File transfer is provided by a service called FTP (which stands for File Transfer Protocol). FTP, which is a basic service available on all Internet computers, provides broad access to network data resources. Unfortunately, FTP sometimes uses network capacity unnecessarily because FTP always transmits a complete file even if only a small extract is needed.

Another network service of interest to data-intensive researchers is Network File System (NFS) which enables transparent remote file access. With NFS, researchers can access a remote file or directory as if it were on his or her local computer. Thus, NFS overcomes the major shortcoming of FTP (NFS does not transmit an entire file). Unfortunately, NFS is not yet as widely available as FTP.

While FTP and NFS enable network access to data stored on almost all Internet computers, they do not automatically convert binary coded numeric information between different computer models. Thus, FTP does not support totally transparent data sharing. Some database software vendors have solved this problem (see section IV).

FTP is a nearly universal tool for providing network access to data resources. But do the Internet and local campus networks really work fast enough to support network access to large research databases? Or does network traffic and the existence of "weak links" (e.g. low-speed regional network connections that become a data transmission bottleneck) reduce performance to the point that transmitting large research databases is infeasible?

To get a realistic assessment of network performance, an informal test was carried out with the help of Jim Jacobs of the University of California at San Diego. A file containing roughly 10 Megabytes (an extract of the General Social Survey) was repeatedly transferred, in roughly two hour intervals, through the national Internet (from San Diego to Syracuse) and through the Syracuse University Internet (between workstations in separate buildings) on Wednesday, May 9, 1990. The data transfer rate was recorded each time the file was transferred.

The results of the test, summarized in Table 1, indicate that campus networks which run at ethernet speeds, such as the Syracus University Internet, are indeed suitable for large data file transmission, while long distance transmission via Internet is limited — for larger files magnetic tape is still an attractive method. Of course within a couple of years the Internet's national backbone network and many of the regional networks will be seeing a 24-fold increase in performance. The day may soon come when magnetic tapes are used to transfer data only in rare instances.

Mass Storage

One of the most striking developments over the past few years in mass storage technology is the proliferation of high capacity storage products. Not too many years ago, if one had a large research database there were only two realistic storage alternatives: it could go on mainframe mag tape or, with a little help from the mainframe systems folks, it might be put up on mainframe magnetic disk. Today there are a slew of other alternatives suitable for a range of large research database applications —

TABLE 1. Results of Informal Internet Data Transfer Test

(EDT)	Transfer Rate	ElapsedTime	Transfer Rate	Elapsed Time
6:33AM	12 kbytes/sec.	14 min.	76 kbytes/sec.	2 min.11 sec.
9:18AM	11 kbytes/sec.	15 min.	74 kbytes/sec.	2 min.15 sec.
11:27AM	10 kbytes/sec.	17 min.	77 kbytes/sec.	2 min. 9 sec.
1:21PM	8 kbytes/sec.	21 min.	71 kbytes/sec.	2 min.20 sec.
3:09PM	8.6 kbytes/sec.	19 min.	76 kbytes/sec.	2 min.11 sec.
5:03PM	8.8 kbytes/sec.	19 min.	76 kbytes/sec.	2 min.11 sec.
7:47PM	8.6 kbytes/sec.	19 min.	71 kbytes/sec.	2 min.20 sec.
9:08PM	7.8 kbytes/sec.	21 min.	75 kbytes/sec.	2 min.13 sec.
11:24PM	11 kbytes/sec.	15 min.	71 kbytes/sec.	2 min.20 sec.

from 700 megabyte hard disks which cost a few thousand dollars and can attach directly to a PC or workstation to mainframe attached, terabyte capacity (i.e. a thousand gigabytes) optical jukeboxes. See the Appendix for descriptions of a wide variety of storage products.

With networks enabling high speed access to a heterogenous mix of computers, one is no longer constrained to a particular computer platform when evaluating mass storage alternatives. Other characteristics of the application — e.g. required capacity, expected frequency of access, and life expectancy of the data — can be matched against the storage technologies available for a variety of platforms.

While there are hundreds of options available for storing large amounts of research data, almost all storage products store data on one of three basic media: magnetic tape, magnetic disk, or optical disk.

Magnetic Tape

Magnetic tape is the storage technology upon which many research data libraries were built, and it remains a dominant research data storage medium on many campuses.

Universities made the rather substantial investment to provide tape access some time ago, so researchers have ready access to central tape storage facilities and tape drives. Magnetic tapes remain an attractive medium for storing large databases because the main cost is that of new tapes, which are relatively inexpensive. A standard reel of 9-track tape, which holds up to 180 megabytes of data, can be purchased for around $15. IBM mainframe 3480 cartridge tapes, which hold slightly more, cost under $10.

Tapes are often used to transport databases between institutions. Tapes can be packed and shipped overnight, and standard tape formats exist which can be read and written on every major university campus in this country.

One of the major problems with magnetic tape is slow data access — the operator intervention required to mount a tape and the sequential processing of magnetic tape results in access time measured in minutes.

Another problem with magnetic tape is limited archival life. Tape is a relatively fragile storage media, with an average archival life of somewhere around five years. Those charged with maintaining access to data on tape for extended periods must periodically "exercise" each tape to ensure readability and prevent print-through.

Compact, high capacity tape products commonly used to back up workstation and minicomputer hard disks have potential as storage media for research databases. 8mm tape store up to 2.3 gigabytes of data in compact cartridge, and an 8mm Exabyte drive (Exabyte is the only 8mm drive manufacturer although several companies sell Exabyte drives) runs $4,000. 4mm tapes, also known as Digital Audio tapes (DAT), are compact cartridges (smaller than a pack of cigarettes) which store 1.3 gigabytes of data. Although 4mm tape drives do not have the installed base of the 8mm drives, buyers are attracted to DAT because drives are made by more than one vendor. Significant increases in the capacity of both 4mm and 8mm tapes are expected.

Magnetic Disk

The basics of magnetic disk technology have not changed significantly in twenty years, but continual improvements have resulted in steady increases in capacity and performance and a steady decrease in cost per megabyte. Magnetic disks are the obvious choice if high performance on-line access to research databases is required.

Hard disk are now available which attach directly to PCs or workstations and hold around 700 megabytes of data. They can be purchased for as low as $2,500. For those with greater appetite for local storage, several drives can be daisy-chained together to provide access to several gigabytes of on-line storage.

Network servers are computers that are dedicated to the task of data access for network client computers. They typically have attached disks that are faster than those attached to individual desktop computers. Network server performance is likely to be boosted in the near future with the introduction of RAID (Redundant Array of Inexpensive Disks) technology. RAID servers will achieve much faster transfer rates by transmitting data in parallel, using multiple disks and read/write heads.

Mainframe disk drives, also known as Direct Access Storage Devices (DASD), currently offer the best overall performance. Mainframe DASD, which can provide on-line access to hundreds gigabytes to hundreds of mainframe users, are typically used for demanding time-sharing and transaction processing applications.

Optical Disk

There are three basic optical technologies on the market today: CD-ROM which is primarily a publishing media with data disks created and distributed by information providers, WORM disks which enable a single write followed by unlimited read access, and erasable magneto-optical disks which allow unlimited read/write access. Digital paper is a newer ultra-high density optical technology which is just becoming available in commercial products.

CD-ROM, WORM, and magneto-optical are very dense optical disk storage technologies, with the disks typically available as removable cartridges or platters. Optical disks are slower than magnetic disks, but the drives are not as susceptible to head crashes and other malfunctions. Optical disks also have a lengthy archival life with some vendors claiming up to 30 years.

CD-ROM disks, developed originally for the audio industry, are 4.7" disks that hold roughly 600 megabytes of data. Data is formatted and written on a CD-ROM master disk (a process called mastering) which acts as a template for "stamping" copies. CD-ROM disks can be mastered for one to two thousand dollars; disks stamped from the master run $2 to $3 per disk.

CD-ROM is popular medium for distributing textual and bibliographic databases and is beginning to catch on as a medium for distributing research data files, a development boosted by the Census Bureau's decision to distribute much of the 1990 data on CD-ROM. Its main attraction is on-line access to large databases from desktop workstations.

In many instances, CD-ROM is a good alternative to dial-up access to expensive information services such as Dialog. However, the general suitability of CD-ROM for research database access is questionable. CD-ROM is relatively slow when compared to other on-line storage media (e.g. hard disks and WORM disks) so it is far from ideal for supporting multi-user on-line access to research databases (though several CD-ROM network server packages are on the market). Furthermore, the computers which many CD-ROM distributors target, often lack the computing resources to effectively handle the analysis files that are typically derived from the large master files distributed on CD-ROM.

WORM (Write Once Read Many) is a high capacity, locally written storage media with a lengthy archival life. WORM is faster than CD-ROM, although WORM drives are not nearly as fast as most magnetic drives.

One of the striking things about WORM technology is the wide range of WORM products. Unlike CD-ROM, WORM is available in several sizes and configurations — from 5-1/4" disks which hold hundreds of megabytes to 14" platters with an 8.2 gigabyte capacity. Most WORK drives act like magnetic disks, although mainframe attached drives typically emulate a tape drive. WORM is available as a single drive removable cartridge drive in some products and in jukebox configurations in others. WORM products are available for the complete range of computer platforms — from PCs to mainframes.

Magneto-optical disk, a recently introduced optical technology, is a high capacity, fully erasable, removable storage media. It shares many of the properties of WORM drives (similar capacity to store data, comparable performance, long archival life), but it is not available in such a wide range of products — 5-1/4" disks which can store several hundred megabytes are the norm.

Digital paper is an ultra-high write-once optical media which can be produced in large sheets and reels. Only one product using digital paper is on the market right now (the Creo 1003 tape drive which stores a terabyte of data on a single reel of tape); and one product that was being planned has been dropped (a Bernoulli drive based on optical paper). The future of digital paper is unclear, but if the technology takes off, it could become the storage media of the future.

Database Sofware

Database software packages enhance database access by relieving the end-user from the burden of knowing the physical characteristics of each variable, for example where each variable is physically located, how long each variable is, and so forth. Once a raw file has been read into a database package, end-users can simply access variables by name, usually with some a flexible, user-friendly query language. This is an excellent approach for master files which are used by groups of researchers.

Database access can also be enhanced by using the indexing capability built into most database software. A database index is essentially a computerized lookup table that speeds data access, similar to the way the index in the back of a book works. By indexing the variables which are frequently sorted on or used to select extracts, researchers can realize significant time savings.

The ability of some database packages to provide transparent access to remote databases is another way database software can enhance access to research databases. Several database packages (e.g. Ingres and Oracle) advertise remote access to data on different platforms through high speed networks. SAS will soon offer an add-on product, called SAS Connect, which will provide the same capability for SAS datasets.

[1] Paper presented at the 16th Annual Conference of the International Association for Social Science Information Service and Technology (IASSIST), Poughkeepsie, New York, May 30-June 2, 1990.

APPENDIX

Mass Storage Products

Product	Description
CompuAdd ESDI Hard Drive CompuAdd 1-800-627-1967	An inexpensive hard disk drive for PC compatibles with a 680 megabyte capacity. Costis around $3,000 including controller card.
Panasonic LF-5010 Panasonic Communications and Systems Co. 1-800-742-8086	An inexpensive 5-1/4" WORM drive which can attach to MS-DOS, UNIX, Xenix, or VMS computers. Drive can access one side of a dual-sided disk - disk capacity is 896 megabytes, while on-line access is limited to 448 megabytes. Cost of drive and interface kit $3,700 (for PC). A jukebox version which holds 50 disks (total capacity 47 gigabytes) is available for $35,000.
Laserstore Erasable LSE-1000 AT Storage Dimensions 1-408-879-0300	An erasable magneto-optical 5-1/4" disk drive with SCSI adapter. Capacity of the two-sided disk is 920 megabytes with 470 megabytes on-line at any one time.cost of drivee, cables, and software is $8,000. The LSE-1000 AT is is also available in a dual drive model.
PC/Novell file server	A Novell network server can be configured with the Netware software, a 286 PC server, and over a 1 gigabyte of hard disk storage for around $10,000. A 386 PC server running Novell Netware with a 4.6 gigabyte disk capacity is $21,450; a 9.2 gigabyte server will run almost $36,000.
SPARCserver Sun Mircosystems, Inc. 1-800-821-4642	The SPARCserver family are Unix based network servers. A wide range of price/performance is offered -- from servers targeted for small workgroups to high performance servers for large databases and/or large numbers of network clients. A typical low end server would be a SPARCsystem-1 with four gigabyte hard disk storage capacity (disk drives from Hewlett Packard) for around $20,000. A high capacity, high performance server with over 16 gigabytes of disk costs around $200,000, while a server with twice that capacity (the maximum SPARCserver capacity) is over $350,000. [All prices are university discount prices.]
Map Assist Fresh Technology Group 1-602-497-4200	For infrequent Novell netork access to a single CD, Map Assist software ($249), a CD-ROM drive ($600), and a PC server ($2,000) is an inexpensive solution. However, relatively poor performance makes this solution less than adequate when heavy simultaneous access is required.
CD Server CBIS, Inc. 1-404-446-1332	CD Server includes a 286 PC, Toshiba CD-ROM drive, and software for $5,300. CD Server runs on Novell networks and gives large work groups fast access to multiple CD-ROM databases. With CD-Server you can connect up to 14 CD-ROM drives to one dedicated server.
Inspire Alphatronic, Inc. 1-800-229-8686	A magneto-optical, removable, erasable optical disk drive which comes in single drive, dual drive, or jukebox models. Each two-sided disk has a 650 megabyte capacity with 325 megabytes on-line at any one time. A single drive unit costs $13,350, a dual drive unit $19,950, and single to dual upgrade is $7,195. The 25 disk jukebox model run around $49,000 while the 50 disk version is $74,000.

3390 DASD and 3990
Controller
IBM

High performance direct access storage for IBM mainframes.Access speeds and data transfer rates indicate performance which exceeds that of any other product reviewed. These devices can be configured to provide high performace access to hundreds gigabytes of DASD.

Creo 1003 Optical Tape Drive
Creo Products, Inc.
1-604-437-6879

A digital paper tape drive. Each digital paper tape, a 12" reel with 35mm wide digital paper, can store one terabyte(1000 gigabytes) of data. The drive can select any byte on the tape in an average time of 28 seconds, and the drive supports a 3 megabyte per second data transfer rate. The drive has a SCSI interface. It sells for $225,000.

Epoch-1
Epoch Systems, Inc.
1-800-U.S.-EPOCH

A hierarchical storage server that provides on-line storage to vast amounts of data. Main memory, magnetic hard disk, and optical disk provide NFS access to data with frequently used data available in memory or on magnetic disk and infrequently used data automatically moved to optical disk. The server can be configured to provide access to up to a terabyte of data.

System 6800
Kodak
1-800-445-6325

A 14" optical disk system designed for centralized data access on minis and mainframes. Each 14"removable disk holds 8.2 gigabytes. A single drive with SCSI interface costs $47,000. A Pertec tape interface unit ($24,000) or an IBM tape interface unit ($74,000) enables the drive to emulate a tape

Management of Machine-Readable Social Science Information
June 10 - June 14, 1991
9 a.m. - 5 p.m.
ICPSR
Ann Arbor, Michigan

This workshop is designed for individuals whose responsibilities include providing access to social science statistical data files to users of these resources. The objectives of the workshop are to introduce information management, data control, and data librarian procedures and techniques.

The course will cover: "Data Orienteering," providing reference services for machine-readable data, issues of developing collections of machine readable data, and problems which ICPSR Official Representatives face in providing data services on their campuses.

"Data orienteering" consists of some basic skills useful to those who manage or support data libraries or research data services that provide local access to ICPSR data files. The following topics will be discussed:

- understanding data organization (rectangular, hierarchical, and relational) and its importance to statistical packages and secondary data analysis;
- interpreting codebook descriptions of data;
- identifying different data file formats and their applications (e.g., raw data files, OSIRIS dictionary and data files, SPSSx system files, etc.);
- preparing subsets of variables and/or cases using one of themajor statistical packages (e.g., SAS, SPSSx, or OSIRIS); and
- learning how to clean data files.

Exercises will be provided with each topic. Topics included in the discussion of reference and collection development will include:

- how to identify appropriate levels of data service;
- how to categorize your local computer and service environment;
- what books should you have in your collection;
- what skills are needed to provide data services;
- how to cite data files;
- what to do about bibliographic control;
- what the collection development issues are.

The discussion will provide an opportunity for participants and presenters to discuss common problems and challenges of providing data reference service in a library environment.

Presenters:

Diane Geraci
Social Science Librarian
State University of New York at Binghamton
DGERACI@BINGVAXA.BitNet

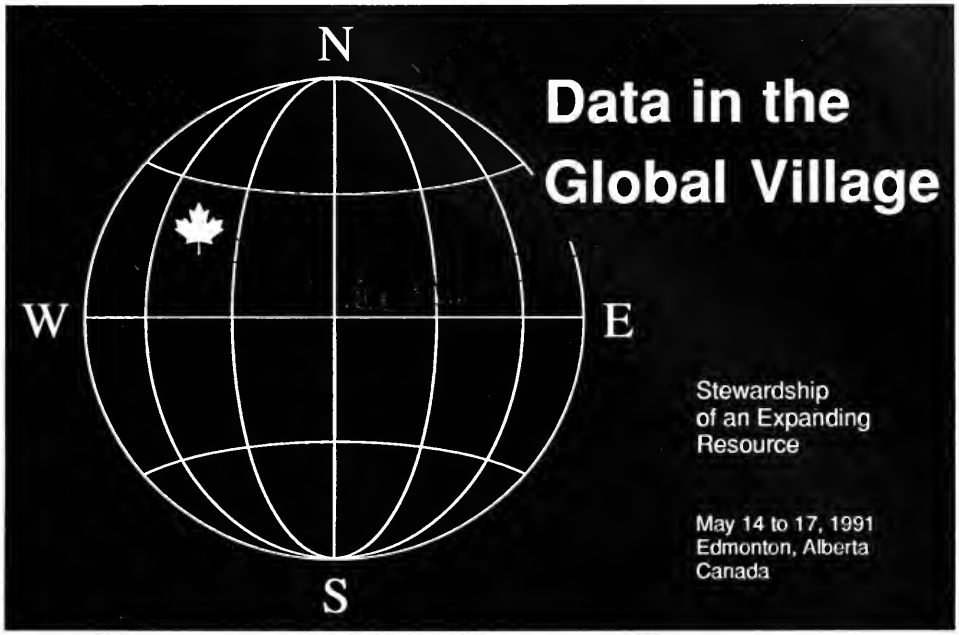
Charles Humphrey
Data Library Coordinator, University Computing Systems
University of Alberta
CHUMPHRE@UALTAVM.BitNet

Jim Jacobs
Data Services Librarian
University of California, San Diego
JAJACOBS@UCSD.EDU

Staff of ICPSR

The fee for the course (tentatively $800), which will be waived for individuals from ICPSR member institutions.

For further information contact:
Henry Heitowit
Program Director
ICPSR Summer Program
P.O. Box 1248
Ann Arbor, MI, 48106
(313) 764-8392



The 17th annual conference of the International Association of Social Science Information Service and Technology will be held at the Hilton hotel in Edmonton, Alberta, Canada, from Tuesday, May 14, through Friday, May 17, 1991. The central conference theme expresses IASSIST members' concern for managing and sharing computer-readable data gathered on a wide range of issues facing our global community. This theme also touches upon the need to care for and preserve an ever-expanding volume of computer-readable data. The conference program features workshops, contributed papers, and roundtable discussions reflecting international viewpoints on these concerns.

About IASSIST

IASSIST brings together individuals from around the world engaged in the acquisition, processing, maintenance, and distribution of computer-readable text and numeric social science data. Founded in 1974, the membership includes data librarians and archivists, information specialists, social scientists, researchers, programmers, planners, and administrators from government and private sectors.

Conference Organizers

Program Committee Chairperson:
Laine Ruus, Data Library Service
University of Toronto,
Toronto, Ontario (416) 978-5589
LAINE@VM.UTCS.UTORONTO.CA

Local Arrangements Coordinator:
Chuck Humphrey, Data Library
University of Alberta Edmonton,
Alberta (403) 492-2741
CHUMPHRE@VM.UCS.UALBERTA.CA

Conference Location

The city of Edmonton lies in rolling parkland, a transitional region etween the open prairies and the Rocky Mountains. The resorts of Banff, Jasper and Lake Louise are within an easy day trip, and in mid-May are still open for skiing. (There will be a post-conference excursion to Banff). Edmonton is the capital city of Alberta and home of the University of Alberta, the Stanley Cup champion Oilers, and of the West Edmonton Mall, the largest indoor shopping and entertainment complex in the world. Come and join us in May for an exciting IASSIST conference!

Special Events

Entertainment activities planned for IASSIST '91 include a period banquet at historic Fort Edmonton and a reception at West Edmonton Mall which will provide an opportunity for conference attendees to 'shop until they drop'.

A post-conference retreat to the spectacular Rocky Mountain resort of Banff is an optional excursion. Because this activity is dependant upon a minimum number of participants, those conference attendees wishing to attend this retreat must register before April 14. To reserve a space, enclose an additional $395 (single occupancy) or $320 (twin occupancy) with your registration. These fees will cover deluxe motor coach transportation to and from Banff, two nights accommodation, and the services of a tour guide.

Transportation
The major airlines with direct flights into the Edmonton International Airport are Air Canada, Canadian Airlines, Northwest Orient, Delta Airlines, and American Airlines. Transportation between the airport and the Edmonton Hilton is available from an airporter service ($9.63) or taxi ($30.00). The Edmonton International Airport is approximately 20 miles from downtown Edmonton.

Accommodations
Rooms have been reserved at the Edmonton Hilton, the conference site, at a special conference rate of $105.00 for a single room and $115.00 for a double. These rates are guaranteed until April 13, 1991 only.

Please contact the Hilton directly to make your reservation before April 13 by mailing the enclosed reservation form or by calling the toll free reservation line (1-800-268-9275). Be sure to identify yourself as an IASSIST conference attendee. Information on alternative accommodations may be obtained from the Local Arrangements Coordinator.

Registration and Fees
If you register before April 14, fees for IASSIST membe s are: $200 for workshops and conference, $175 for conference only, $100 for workshops only, and $100 for one-day attendance. Non-members should add $50 to these fees. After April 14th, a late registration fee of $50 should be added. A new membership in IASSIST is $40 U.S.

To register for the conference, workshops, and Banff post-conference excursion, please return the enclosed registration form along with your payment.

Useful Telephone Numbers
Local Arrangements:
(403) 492-5212

Edmonton Hilton:
Toll-free reservations
1-800-268-9275

In Edmonton (403) 428-7111

Airporter service:
Courtesy phone

Checklist
- ❑ Hotel reservation
- ❑ Airline reservation
- ❑ Conference registration mailed
- ❑ Banff excursion registration

IASSIST '91 CONFERENCE AND WORKSHOPS

Tuesday, May 14, 1991

9:00 -12:30 **Starting a Data Library**

Coordinator: Ilona Einowski

Description: Experienced data librarians will describe how to start a new data library service. Examples of the various activities involved in organizing and operating a data library will be presented.

Living with UNIX ©

Coordinator: Jim Jacobs

Description: Many data libraries are moving to Unix environments. This workshop will provide a basic introduction to Unix and offer a hands-on lab addressing such topics as general Unix utilities, Unix text utilities, varieties of Unix platforms, networking, tape handling, security, and portability.

2:00 -5:00 **Financial Time Series**

Coordinator: Walter Piovesan

Description: Description: With increasing use of financial databases, the staff of data libraries are having to become more familiar with these products. This workshop introduces two such databases, CRSP and COMPUSTAT, and will cover the content of these databases as well as the ways in which users select data from them.

Using Interactive Graphics and Statistical Data in the Classroom

Coordinator: Wendy Watkins

Description: Canada and Norway have projects using interactive graphics that are aimed at putting complex databases into the hands of high school teachers and students. This workshop will introduce Norway's NSDstat+ and Canada's TELICHART software and will allow participants to experiment with both packages during a hands-on session.

Preliminary Program

Wednesday, May 15, 1991

08:00 Registration

09:00 Plenary Session:
The International Global Change Program
Chair: *Peter Burnhill*, University of Edinburgh

10:30 Break

11:00 Concurrent Sessions:
Circumpolar Data Sources
Chair: *Cliff Hickey*, Director, Canadian Circumpolar Institute
Remote Access and New User Services
Chair: *Laura Guy*, University of Wisconsin, Madison
New Data Description Systems
Chair: *Patricia Vanderberg*, University of California, Berkeley

12:30 Lunch Break

02:30 Concurrent Sessions:
Non-Census Microdata
Chair: *Doug Norris*, Statistics Canada
Evaluation and Appraisal of Machine-Readable Data Files
Chair: *Carolyn Geda*, ICPSR
Text File Issues
Chair: *Diane Geraci*, SUNY at Binghampton

06:00 Banquet at Fort Edmonton

Thursday, May 16, 1991

08:00 Registration

09:00 Plenary Session:
Data in the Archival Village
Chair: Terry Cook, National Archives, United States

10:30 Break

11:00 Concurrent Sessions:
The Data Challenge to Librarians
Chair: Judith Rowe, Princeton University
Special Data Collections on Global Change
Chair: JoAnn Dionne, Yale University
Archival Standards of Magnetic Media
Chair: Tom Brown, National Archives, United States

12:30 **Round Table Lunches**

02:30 Concurrent Sessions:
New Software and Hardware Applications for Data Libraries
Chair: Martin Pawlocki, UCLA
Data from the Source: The Data Producers
Chair: TBA
Birthing Pains of Library Data Services
Chair: Linda Langschied, Rutgers University

06:00 **Shop 'Til You Drop at the West Edmonton Mall**

Friday, May 17, 1991

08:00 Registration

09:00 Plenary Session:
International Census Year
Chair: Craig McKie, Statistics Canada

10:30 Break

11:00 Concurrent Sessions:
Electronic Products in Depository Library Programs
Chair: Wendy Watkins, Statistics Canada
International Public Use Microdata Files
Chair: Chuck Humphrey, University of Alberta
Designing and Producing Your Own CD-Rom
Chair: Doug Link, University of Western Ontario

12:30 Lunch and IASSIST Annual General Meeting

02:30 Concurrent Sessions:
Historical Census Data
Chair: Laura Bartolo, Kent State University
Moving Toward Distributed Data Libraries
Chair: Jim Jacobs, University of California, San Diego
Data and Copyright©
Chair: Sarah Cox-Byrne, Vassar College

04:00 IASSIST Nostalgia and Conference Wrapup

06:00 Departures for Banff

IASSIST '91 Registration Form

IASSIST '91 Conference and Workshops May 14-17, 1991 The Hilton, Edmonton, Alberta, Canada

To avoid late fees, please return this form by April 14

Name	
Title	
Affiliation	
Address	
Phone	
E-Mail	

I wish to register for the following:

Member / non-member

- ❑ Workshop(s) only $100 / $150
- ❑ Conference only $175 / $225
- ❑ One day attendance $100 / $150
- ❑ Conference and workshop(s) $200 / $250
- ❑ New IASSIST membership $45
- ❑ After April 14, add a $50 late fee

Please indicate which workshops you wish to attend:

- ❑ _____
- ❑ _____

I will be joining the post-conference excursion to Banff:

- ❑ Single occupancy add $395
- ❑ Twin occupancy add $320

Please make cheque or money order payable to IASSIST (in Canadian dollars) and mail with registration form to:

Chuck Humphrey, IASSIST
352 General Services Bldg.
University of Alberta
Edmonton, Alberta, Canada T6G 2H1

ASSOCIATION FOR HISTORY AND COMPUTING 6TH INTERNATIONAL CONFERENCE ODENSE, DENMARK, AUGUST 28-30, 1991

The Association for History and Computing is an international organization which aims to promote and develop interest in the use of computers in all types of historical study at every level, in both teaching and research. The 6th international conference of the AHC in 1991 will be held in Odense, Denmark, with the Danish Data Archives as the organizing institution. The previous conferences were held in London twice, in Cologne, Bordeaux, and Montpellier.

Themes

A number of workshops on methodological questions will be held in the spring of 1991. The topics of the workshops include text encoding, model solutions for historical problems, high-tech history in the East and West, and image processing. The participants of these workshops will present their results with a view to further discussion in workshop sessions at the conference. It would be natural, as the conference takes place in Scandinavia, that demographic studies and large data collections should be central issues. And, as has become a tradition at the previous conferences, we expect to see and hear presentations of what is happening in the extensive field of history and computing. Among the expected topics are:

Standardization and exchange of machine readable data in the historical disciplines
Data analysis and presentation
Event history analysis
Text analysis
Simulation and modelling
Computer-aided teaching
Social and economic history
Quantitative methods

Other activities at the conference

The General Assembly of the association will take place during the conference. Furthermore, the numerous special working groups of the AHC will, of course, take this opportunity to get together and continue their activities. Among the working groups are some dealing with topics such as the intellectual property of source data banks, archiving and the exchange of data.

Call for papers

Papers are invited on all aspects of computing in history, on substantial subjects as well as on methodological questions. Proposals for papers or demonstrations should be sent to H. J. Marker, Danish Data Archives, Munkebjergvaenget 48, DK-5230 Odense M, Denmark before April 15, 1991.

Proceedings of the conference

Papers presented at the Odense conference and given to us in machine readable form (WordPerfect or ASCII) will be collected and published after the conference.

Place

The conference will be held at the congress center Hotel H.C. Andersen, Claus Bergs Gade 7, DK-5000 Odense Hotel accommodation will be provided by Odense Kongress Bureau, and on the registration form you will find the prices for single bedrooms in three categories. It goes without saying that if you want a double bedroom, this can be arranged by the bureau.

Odense lies on the island of Funen, which is known for its lovely countryside and many castles and manor houses. On the second day of the conference we have arranged a guided tour to start at 3 p.m. The trip will take us across southern Funen and the island of Taasinge. On the way we will pay a visit to Egeskov Castle, which was built on a complex construction of rammed down oak piles and is rising directly from a lake. It was built in 1550 by the Lord High Constable, Frands Brockenhuus.

Registration Fees

The conference fee, which includes conference activities, reception, lunch and coffee on the 3 days of the conference is DKK 900, which must be paid with a cheque, drawn on a Danish bank with the return of the registration form before June 4, 1991 to Odense Kongres Bureau, AHC-Conf., Rådhuset, DK-5000 Odense C. Please note that there will be no refunds of registration fees after July 27. The price for the excursion includes guided tour in a bus, entrance fee to Egeskov Castle and dinner at Hotel Troense and amounts to DKK 300, which must be paid with the Conference fee.

Program Committee

Peter Denley, Westfield College, London;
Stefan Fogelvik, Stockholms Historiska Databas;
Daniel Greenstein, Glasgow University;
Hans Jørgen Marker, Danish Data Archives;
Jan Oldervol, Universitetet i Tromsø:, Norway;
Kevin Schurer, Cambridge Group;
Josef Smets, Montpellier;
Manfred Thaller, Max Planck Institut für Geschichte, Göttingen.

Organizing Committee

Dorte Banke, Kirsten Lund-Jensen and Hans Jørgen Marker, Danish Data Archives.

Further Information

For further information about the conference, please contact

Hans Jørgen Marker
Danish Data Archives
Munkebjergvaenget 48 DK-5230
Odense M Denmark

Phone +45 66 15 79 20 Fax +45 66 15 83 20
E-Mail (EARN): DDAHM@ NEUVM1

IASSIST

INTERNATIONAL ASSOCIATION FOR SOCIAL SCIENCE INFORMATION SERVICE AND TECHNOLOGY

• • • •

ASSOCIATION INTERNATIONALE POUR LES SERVICES ET TECHNIQUES D'INFORMATION EN SCIENCES SOCIALES

Membership form

The International Association for Social Science Information Services and Technology (**IASSIST**) is an international association of individuals who are engaged in the acquistion, processing, maintenance, and distribution of machine readable text and/or numeric social science data. The membership includes information system specialists, data base librarians or administrators, archivists, researchers, programmers, and managers. Their range of interests encompases hard copy as well as machine readable data.

Paid-up members enjoy voting rights and receive the **IASSIST QUARTERLY**. They also benefit from reduced fees for attendance at regional and international conferences sponsored by **IASSIST.**

Membership fees are:

Regular Membership. $20.00 per calendar year.

Student Membership: $10.00 per calendar year.

Institutional subcriptions to the quarterly are available, but do not confer voting rights or other membership benefits.

Institutional Subcription:

$35.00 per calendar year (includes one volume of the Quarterly)

I would like to become a member of IASSIST. Please see my choice below:

- ☐ $20 Regular Membership
- ☐ $10 Student Membership
- ☐ $35 Institutional Membership

My primary Interests are:

- ☐ Archive Services/Administration
- ☐ Data Processing
- ☐ Data Management
- ☐ Research Applications
- ☐ Other (specify)_____

Please make checks payable to IASSIST and Mail to :

Ms Jackie McGee
Treasurer, IASSIST
% Rand Corporation
1700 Main Street
Santa Monica

Name / title

Institutional Affiliation

Mailing Address

City

Country / zip/ postal code / phone

SCSE - QUARTERLY
ISSR
University of California
Hilgard Avenue
CA 90024-1484

BOOK RATE

SERIALS DEPT.
U. OF N. CAROLINA-CHAPEL HILL
DAVIS LIBRARY
CHAPEL HILL, NC 27599

MAY 17, 18, 19